Orbit Correction Principles

John Carwardine

Reasons for orbit correction

- **Get the beam around the machine**
  - A perfect machine doesn't need this.
  - 'Gross' orbit correction (~100's µm scale)

- **Keep the electron orbit through the center of the focusing magnets**
  - Results in better quality electron beam, and hence better quality x-ray beam.

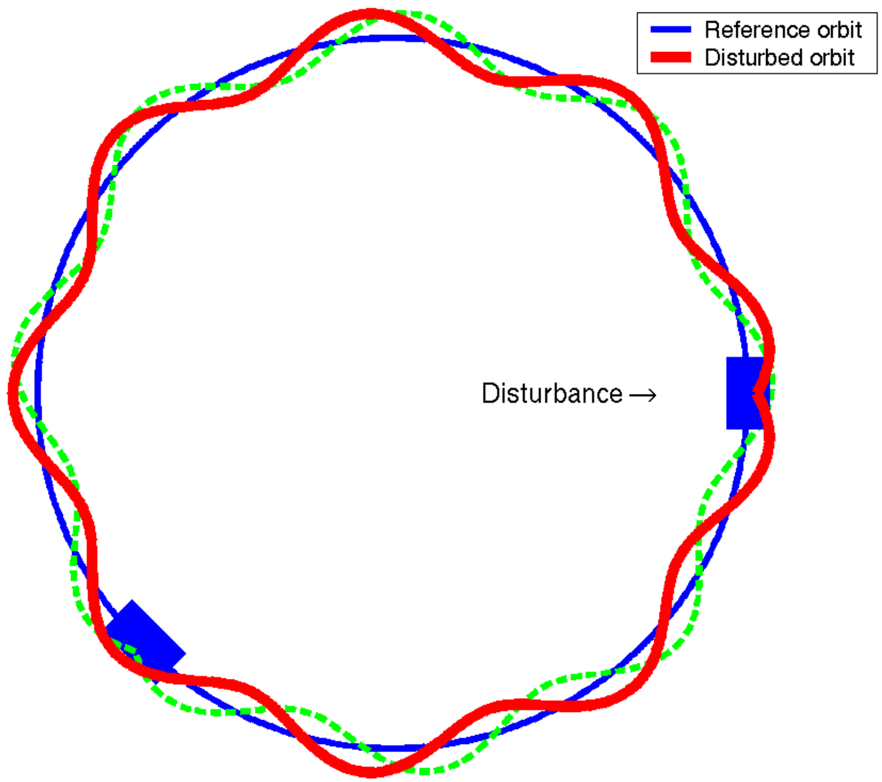
- **Steer x-ray beam away from places it shouldn't be**
  - So we don't melt metal.

- **Steer the x-ray beams at the users' samples**
  - Long lever arm (typically 40-70m) makes it a challenge to get the beam on target.
  - Electron orbit though x-ray source point defines x-ray trajectory.
  - Fine orbit correction (<1's µm scale)

- **Keep the x-ray beam on the users' samples**
  - Long-term effects (drift, etc)
  - Short-term effects (electrical noise, vibration, etc)

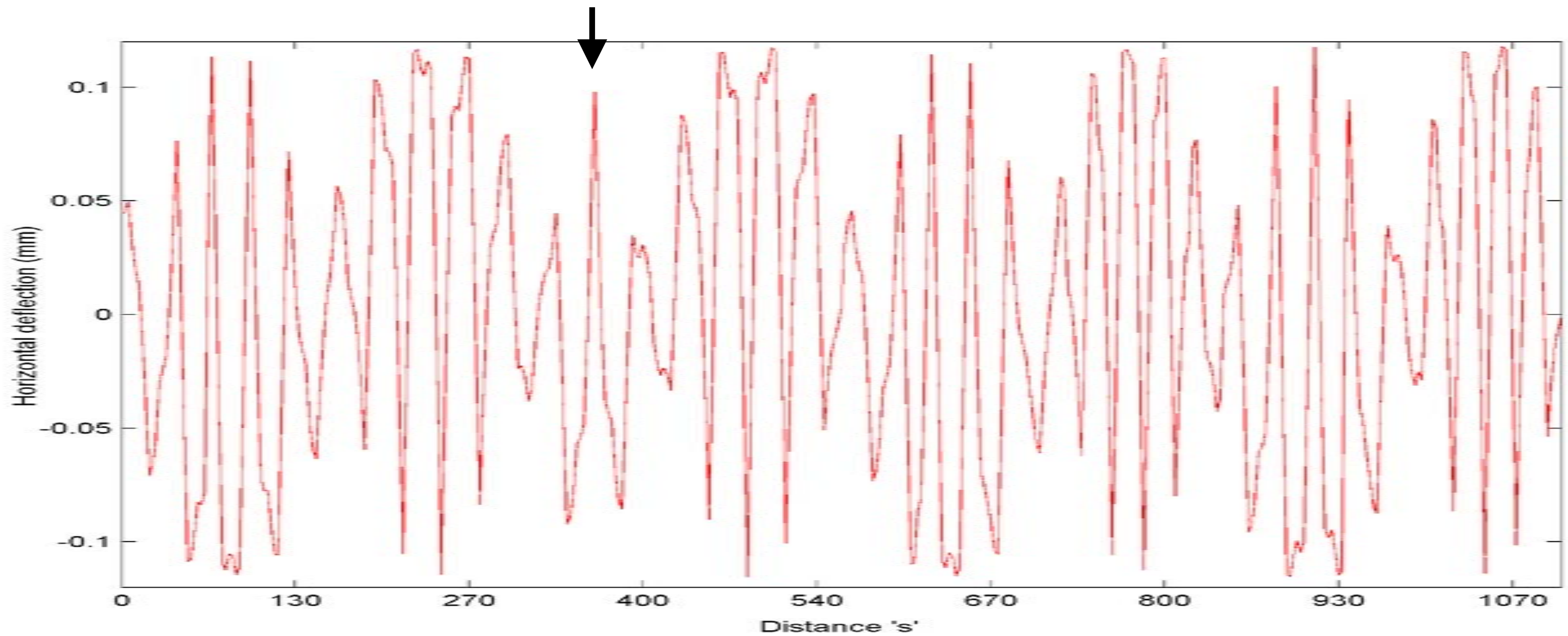
Betatron Oscillations

- Disturbances to the beam produce harmonic betatron motion.
- Regardless of the location of the disturbance, the entire orbit is affected.



Orbit deflection caused by a single source

- The number of oscillations around the entire ring is the betatron tune.
- The phase of the oscillation depends on the location of the source.
- Amplitude of the oscillation depends on the beta function at the source.
- Example below shows the betatron oscillation for the APS horizontal orbit for a single source located around 370m.

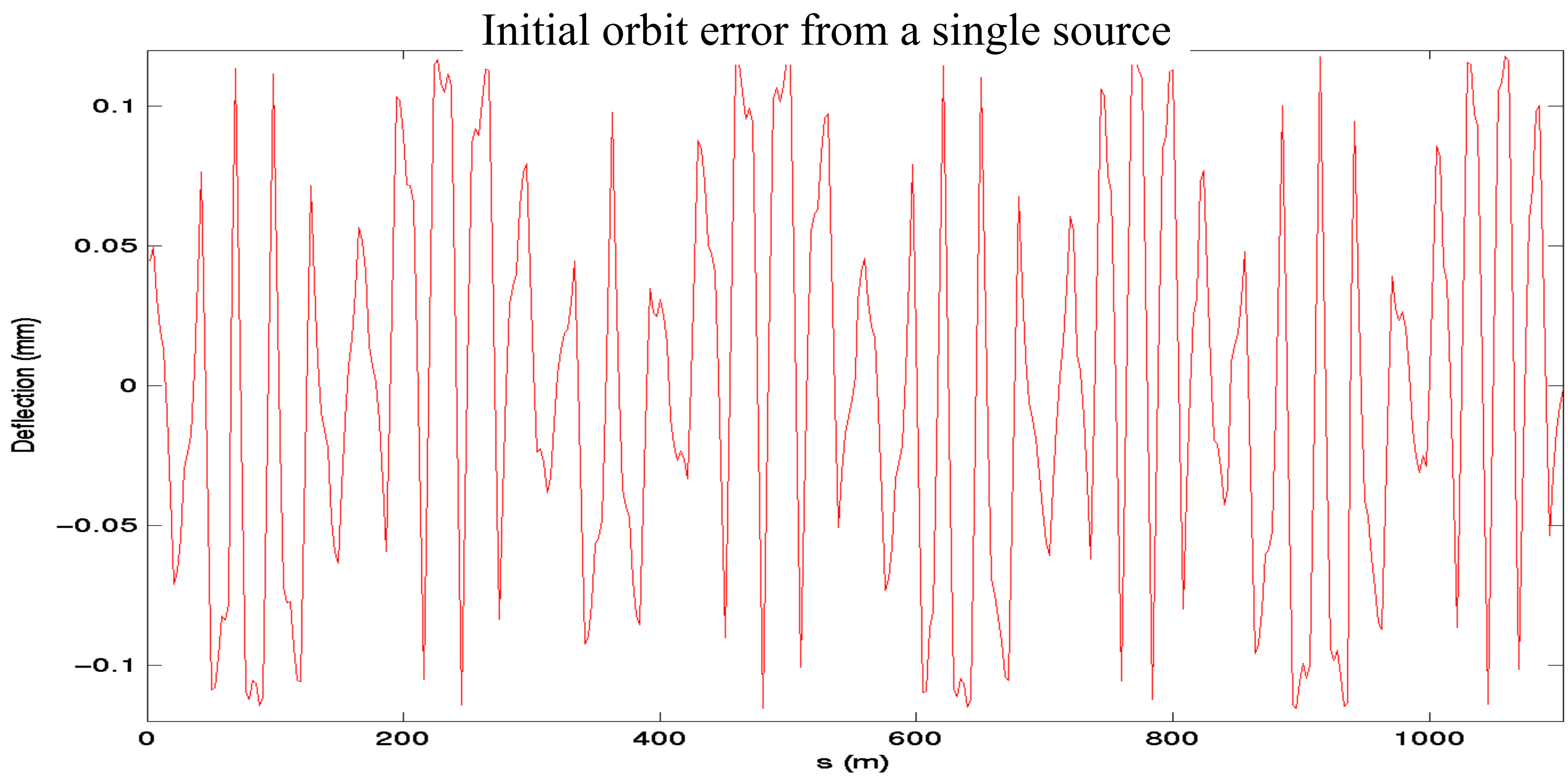


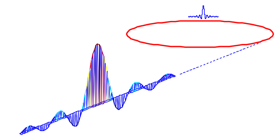
Orbit correction techniques

- Manual correction
  - “Grab a corrector and start tweaking”.
- Local orbit correction
  - Control orbit at specific location without disturbing global orbit.
- Global orbit correction
  - Harmonic correction.
  - SVD.
  - Least-squares.
  - Weighted least-squares.

Fixing betatron oscillation by "grabbing a corrector"

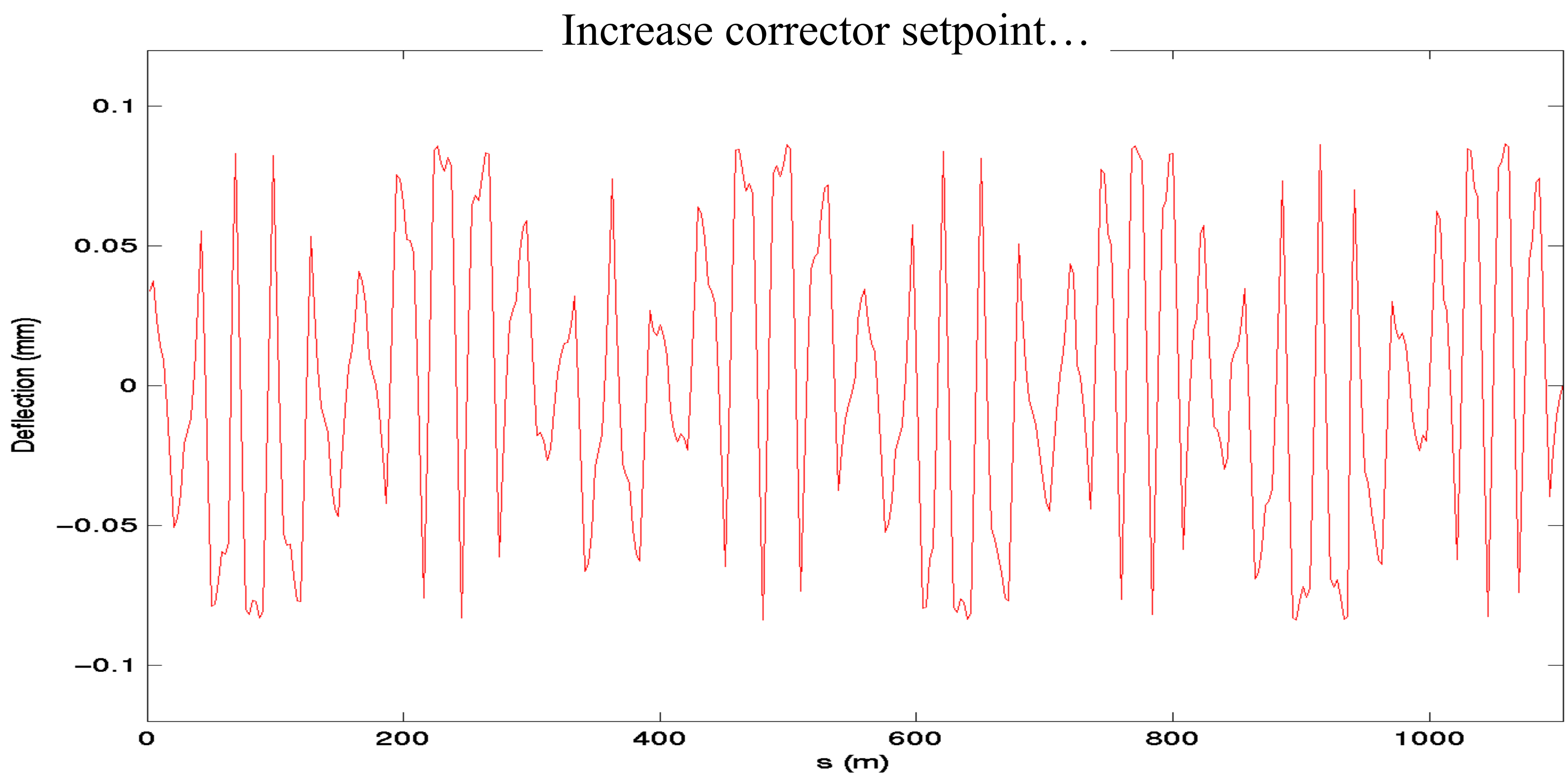
- It is possible to hand correct gross orbit errors caused by strong sources.
- Tweek corrector setpoint until orbit error is minimized.
- Corrector doesn't have to be close to source to make gross improvements





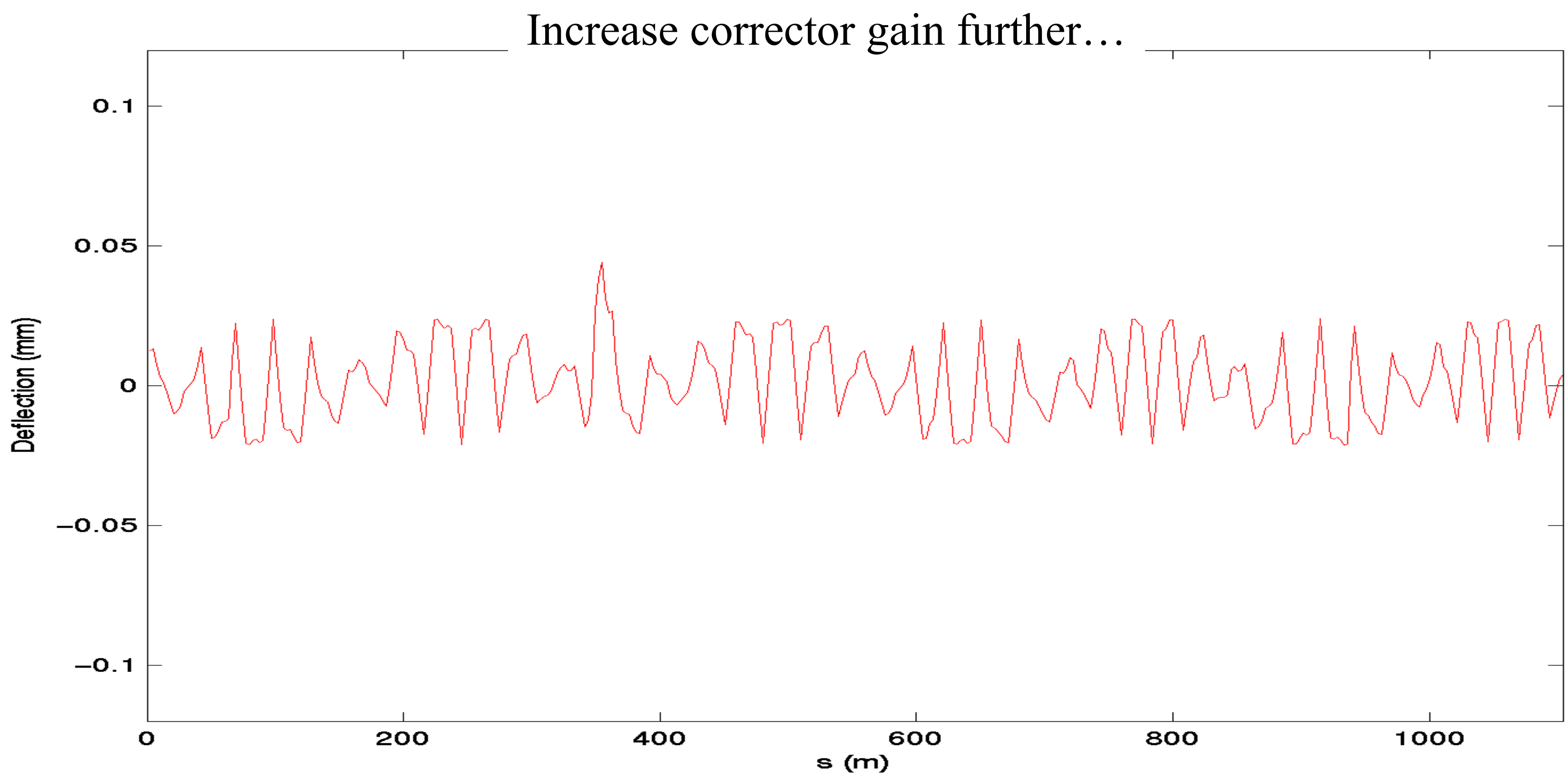
Fixing betatron oscillation by “grabbing a corrector”

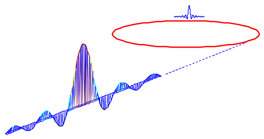
- It is possible to hand correct gross orbit errors caused by strong sources.
- Tweek corrector setpoint until orbit error is minimized
- Corrector doesn’t have to be close to source to make gross improvements



Fixing betatron oscillation by “grabbing a corrector”

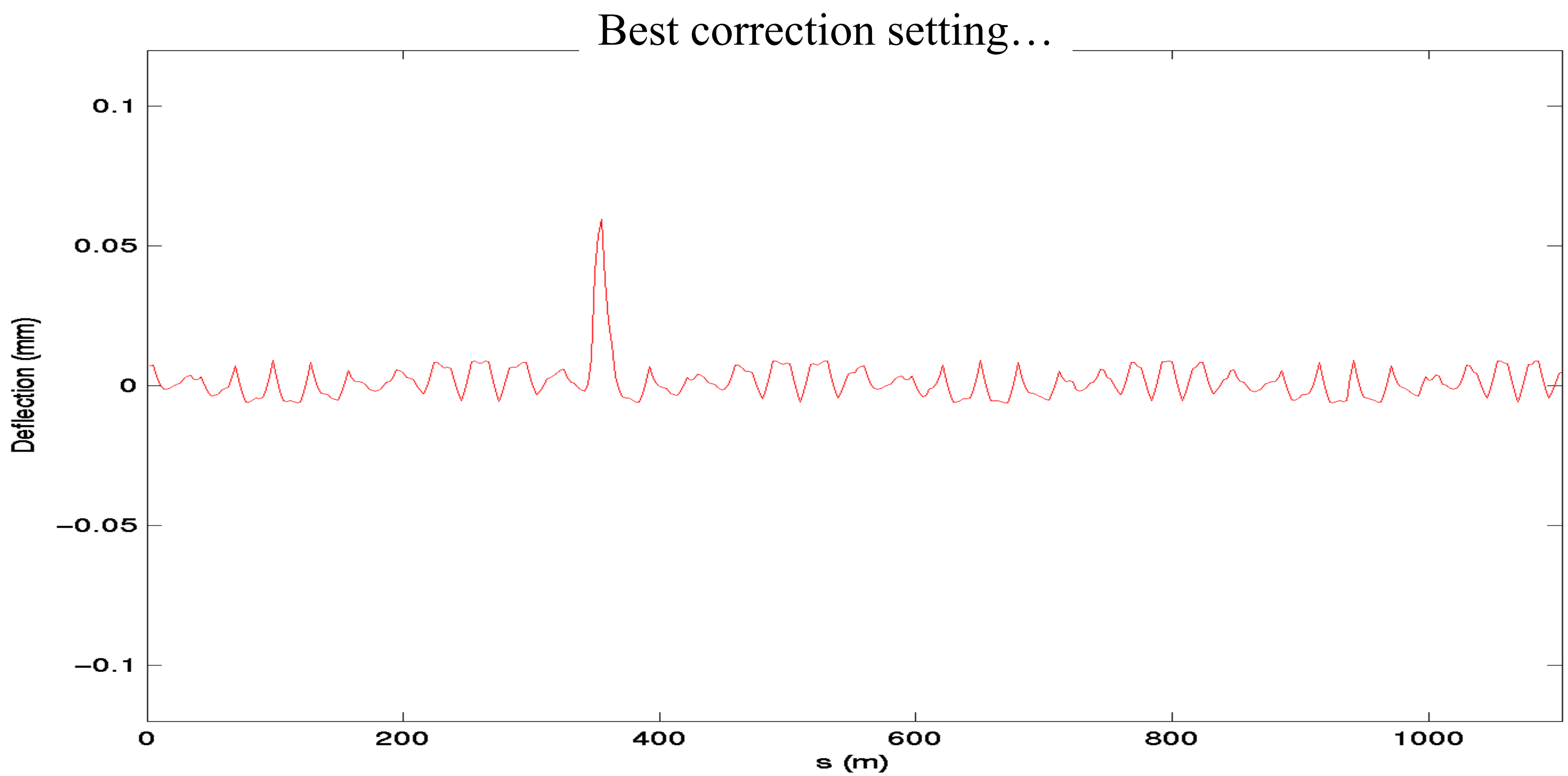
- It is possible to hand correct gross orbit errors caused by strong sources.
- Tweek corrector setpoint until orbit error is minimized
- Corrector doesn’t have to be close to source to make gross improvements

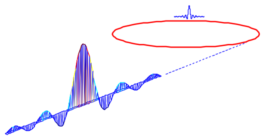




Fixing betatron oscillation by "grabbing a corrector"

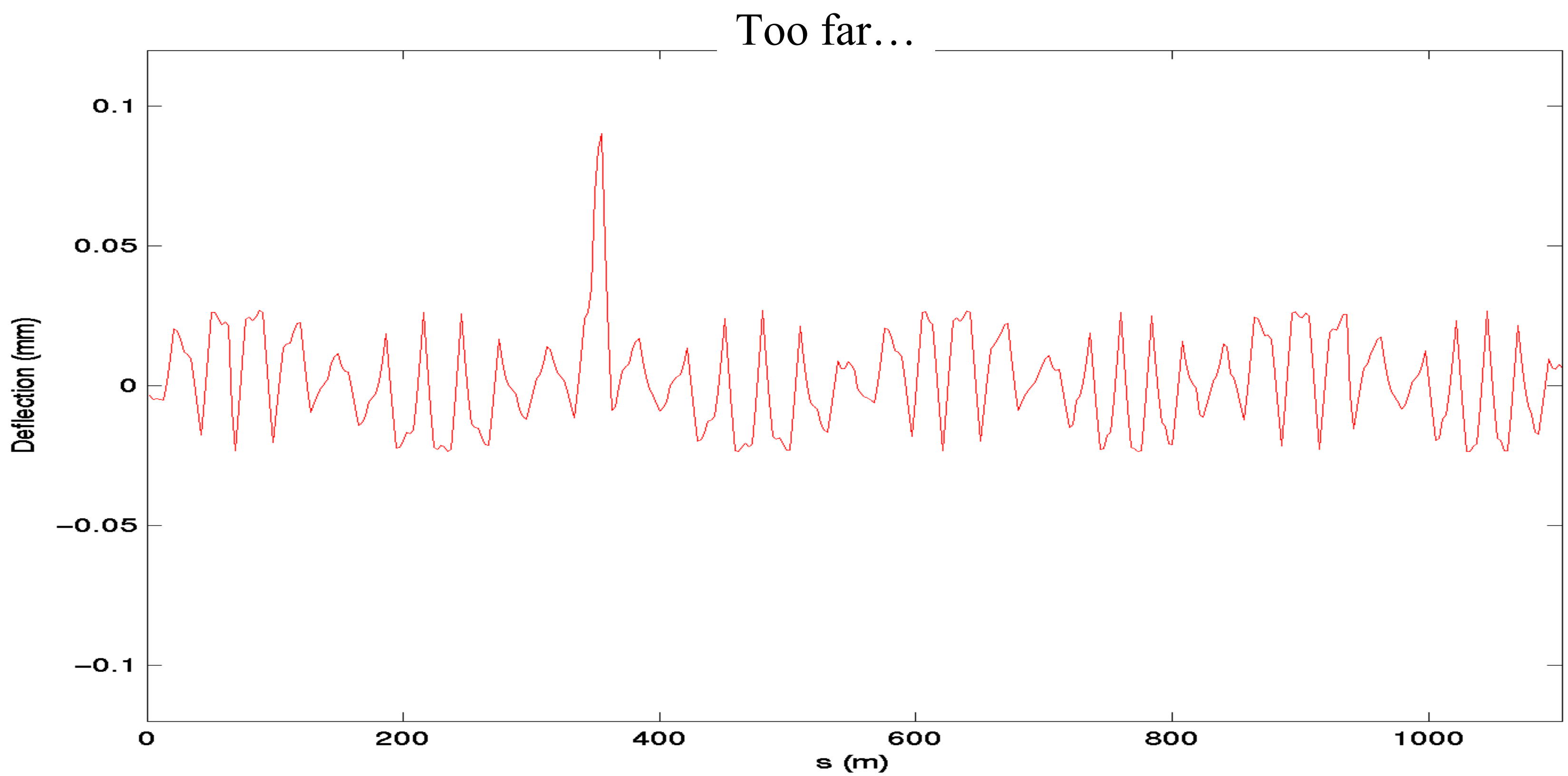
- It is possible to hand correct gross orbit errors caused by strong sources.
- Tweek corrector setpoint until orbit error is minimized
- Corrector doesn't have to be close to source to make gross improvements

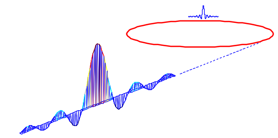




Fixing betatron oscillation by "grabbing a corrector"

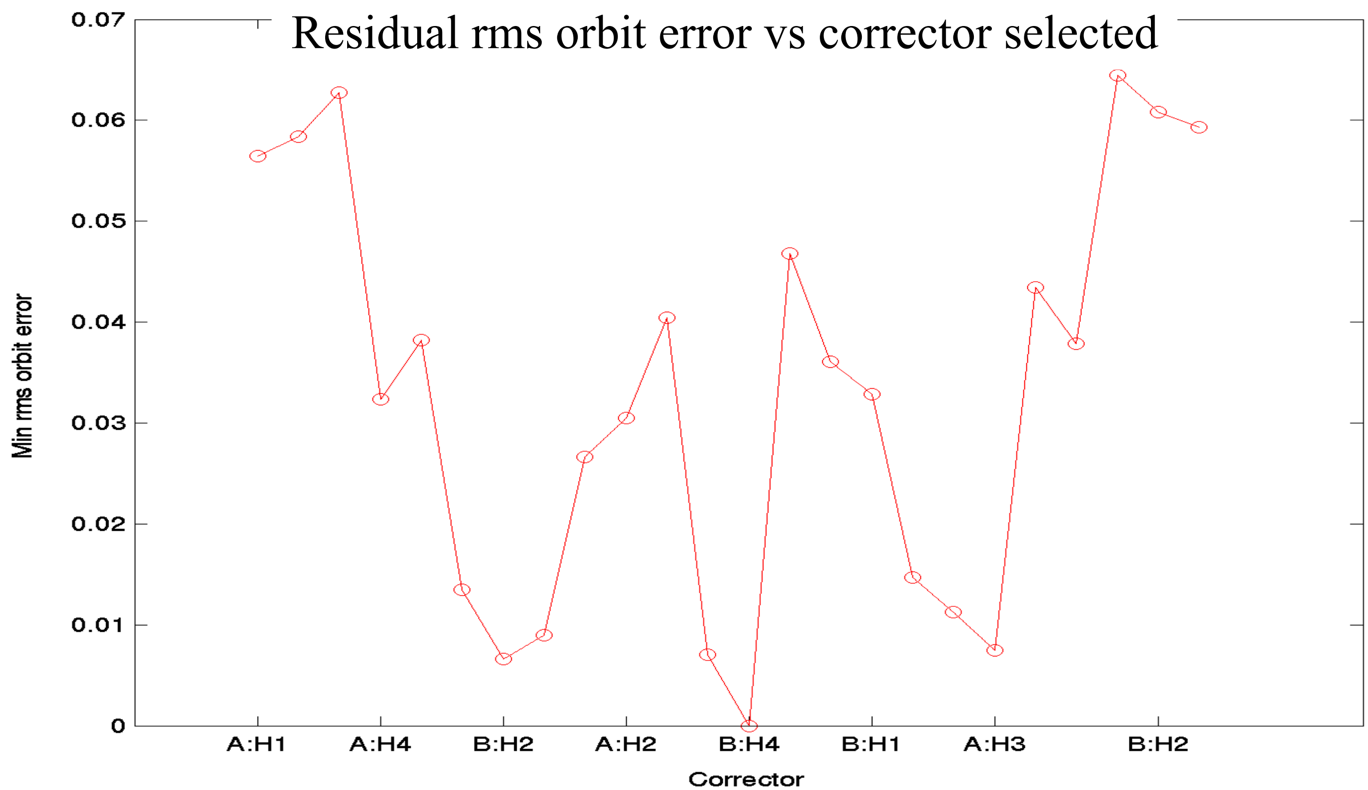
- It is possible to hand correct gross orbit errors caused by strong sources.
- Tweek corrector setpoint until orbit error is minimized
- Corrector doesn't have to be close to source to make gross improvements





Residual orbit vs relative location of corrector

- Although gross of a single source can be made using almost any corrector, the overall effectiveness depends on picking the right corrector.
- Perfect correction is only possible when the corrector is exactly at the same location as the source of the disturbance.



Decompose orbit into two orthogonal components

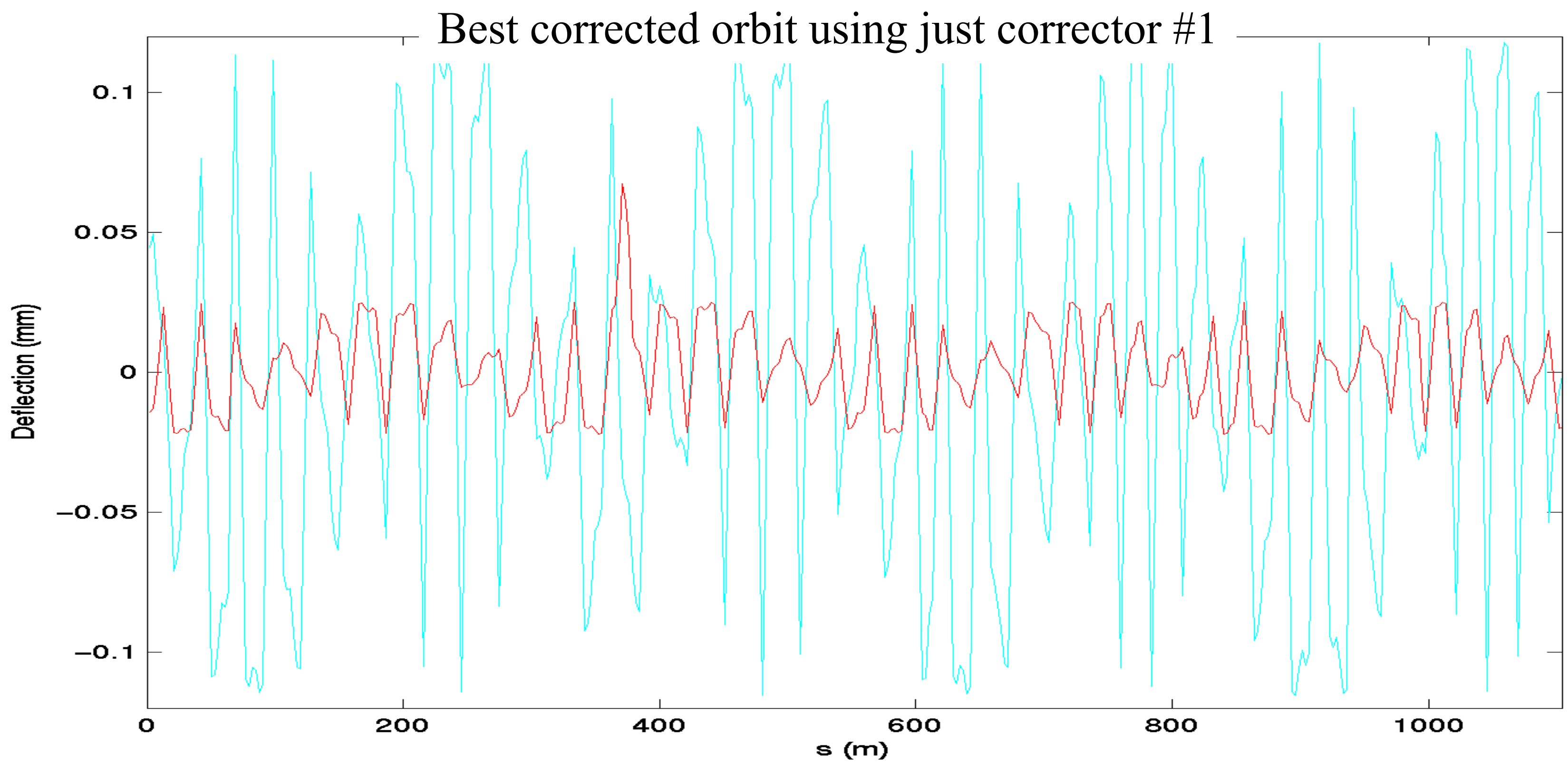
- Knowing that the orbit phase is a function of the corrector location means we can decompose the orbit into two orthogonal components, so the orbit motion starts to look like..

$$x \approx a \sin \omega s + b \cos \omega s$$

- Therefore, we should be able to use two correctors 90 degrees apart to completely cancel the orbit motion…

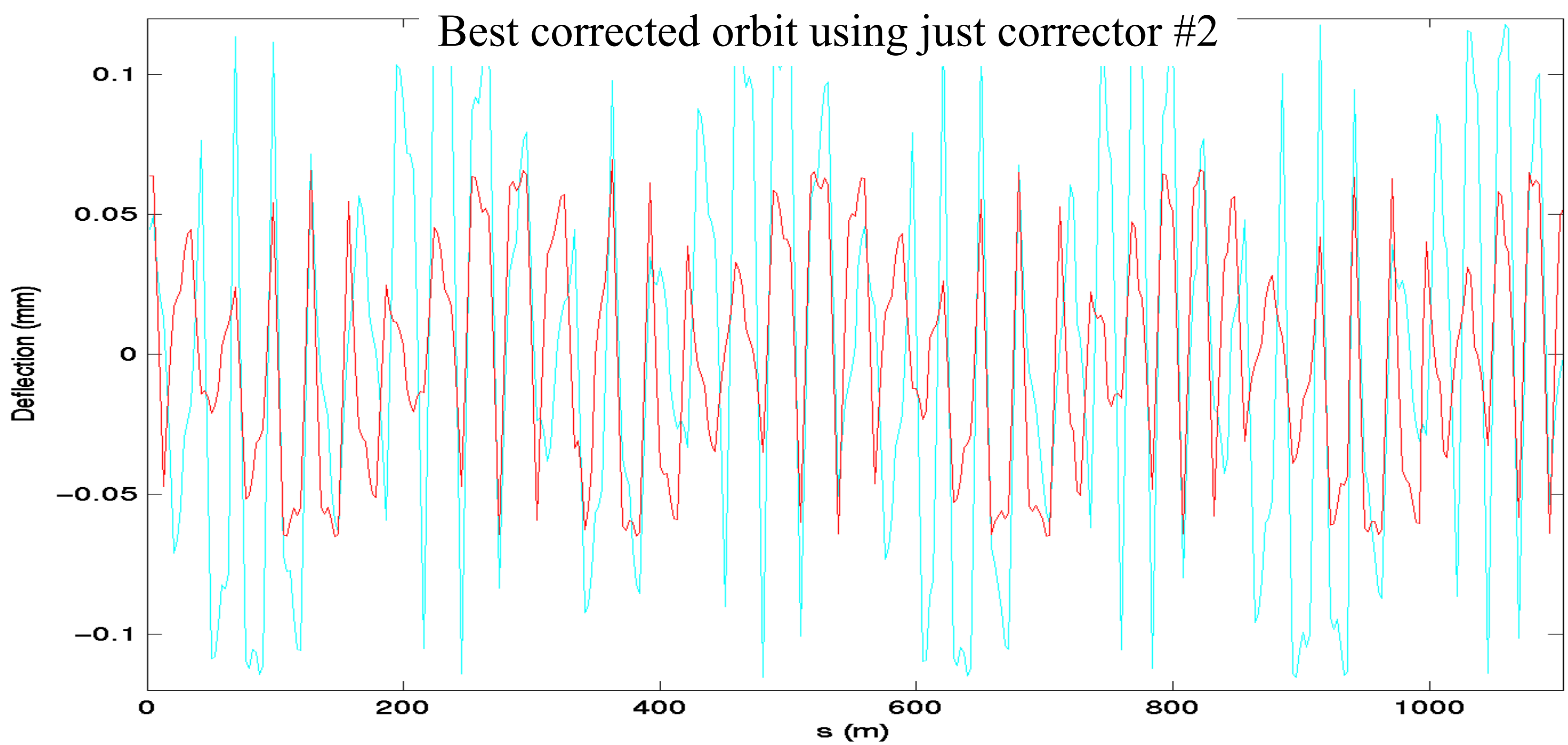
Using two correctors to cancel a single source

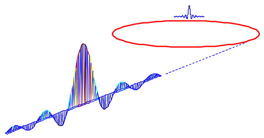
- Adding a second corrector allows the orbit to be fixed globally, leaving only a residual localized disturbance between the two correctors.



Using two correctors to cancel a single source

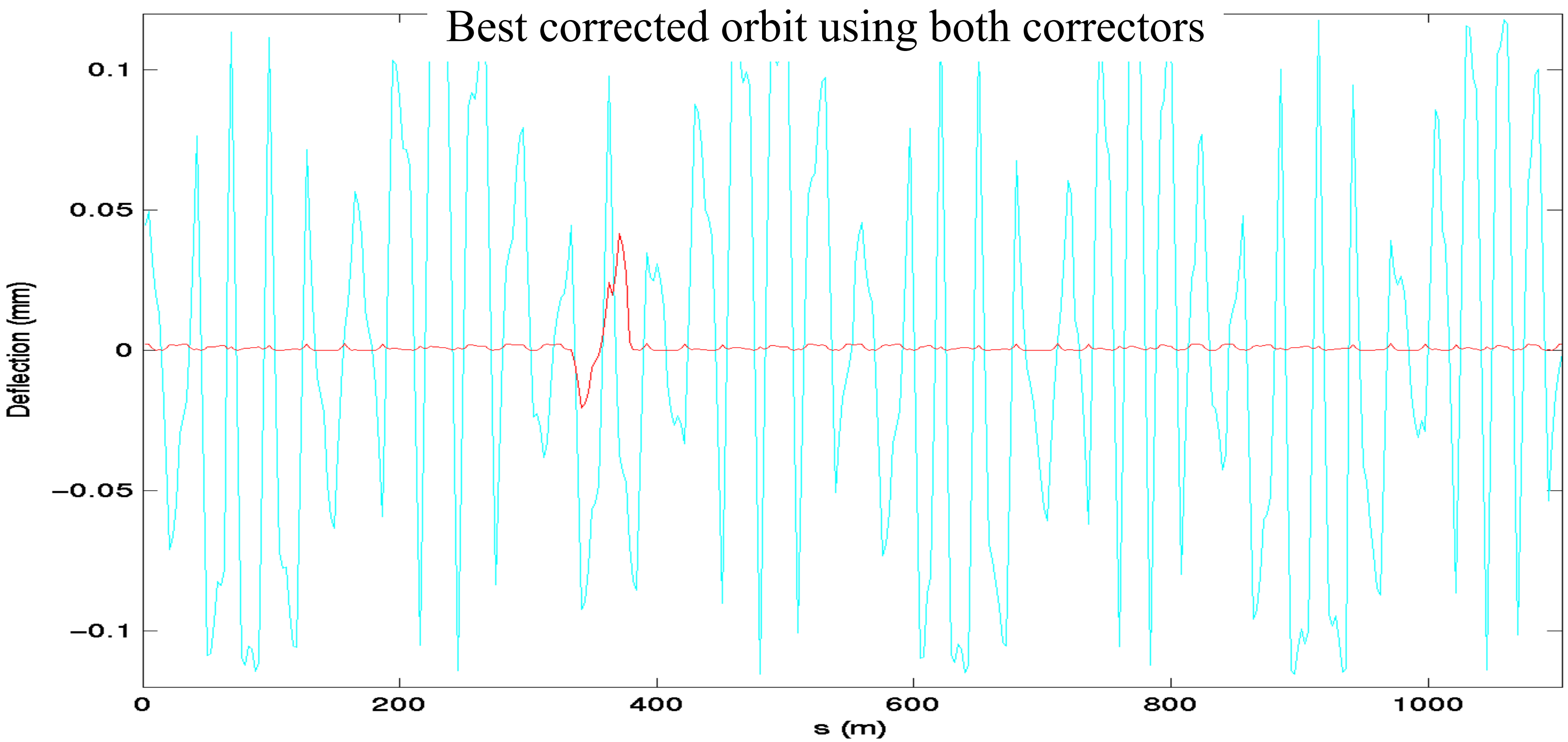
- Adding a second corrector can allow the orbit to be fixed globally, leaving only a residual localized disturbance between the two correctors.

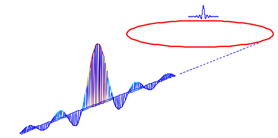




Using two correctors to cancel a single source

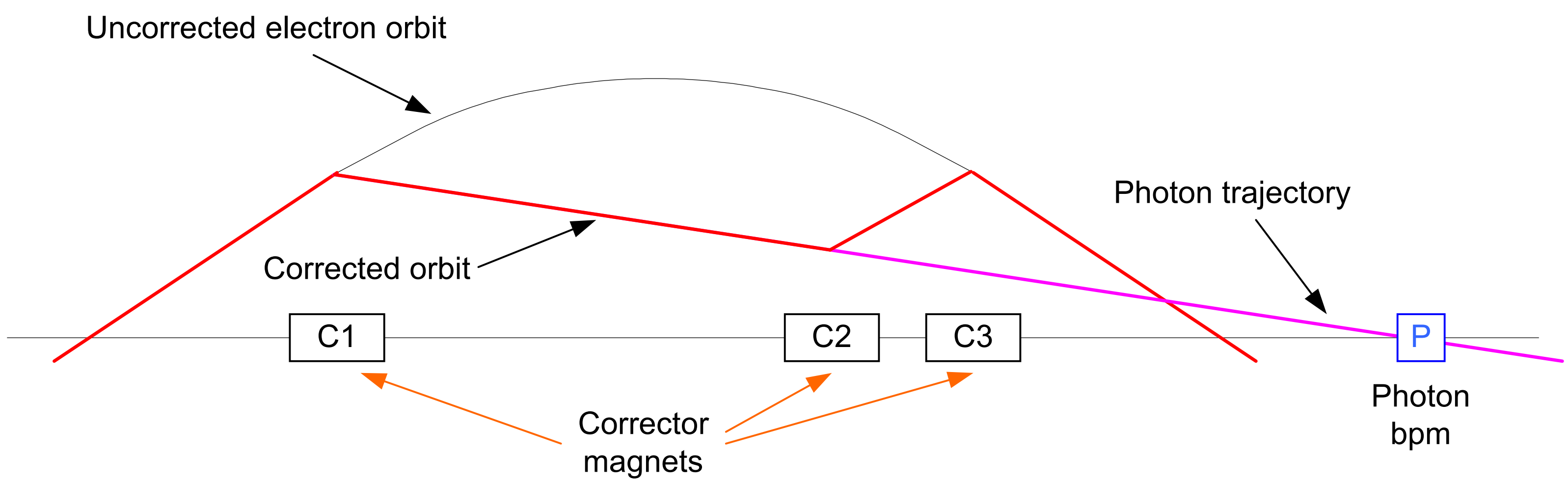
- Adding a second corrector can allow the orbit to be fixed globally, leaving only a residual localized disturbance between the two correctors.



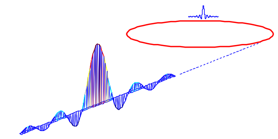


Local orbit correction using three-corrector bump

- A three-corrector local bump can be used to correct local disturbances or to steer the beam through a local source point.
- It provides localized control of the position at a single bpm location, either a photon bpm (as shown), or an rf bpm located inside the bump

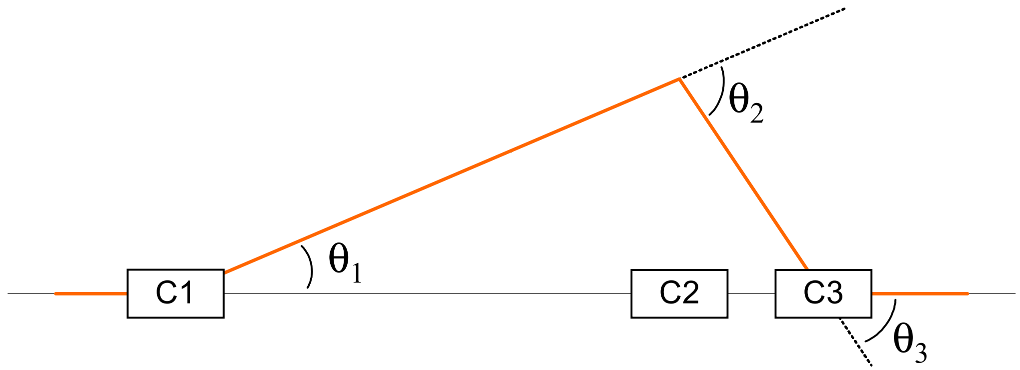


- Requirements for the bump to be confined
  - The orbit position after the last corrector magnet must be unchanged.
  - The orbit angle after the last corrector magnet must be unchanged.



Calculating local bump coefficients

- The conditions are satisfied with fixed ratios of the three corrector strengths that are determined from the beta functions and relative phases at the correctors

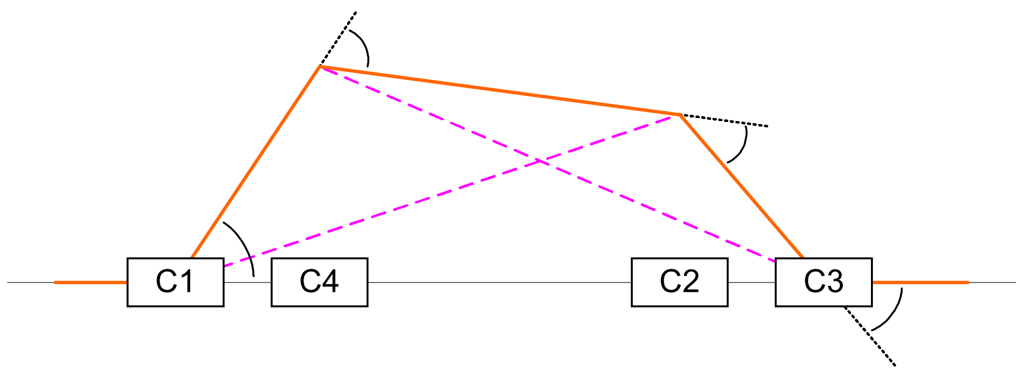


$$\theta_2 = -\theta_1 \left(\frac{\beta_1}{\beta_2}\right)^{\frac{1}{2}} \frac{\sin\phi_{13}}{\sin\phi_{23}} \qquad \theta_3 = \theta_1 \left(\frac{\beta_1}{\beta_3}\right)^{\frac{1}{2}} \frac{\sin\phi_{12}}{\sin\phi_{23}}$$

β_1 = beta function at C_1, etc $\qquad \phi_{12}$ = phase between C_1 & C_2, etc

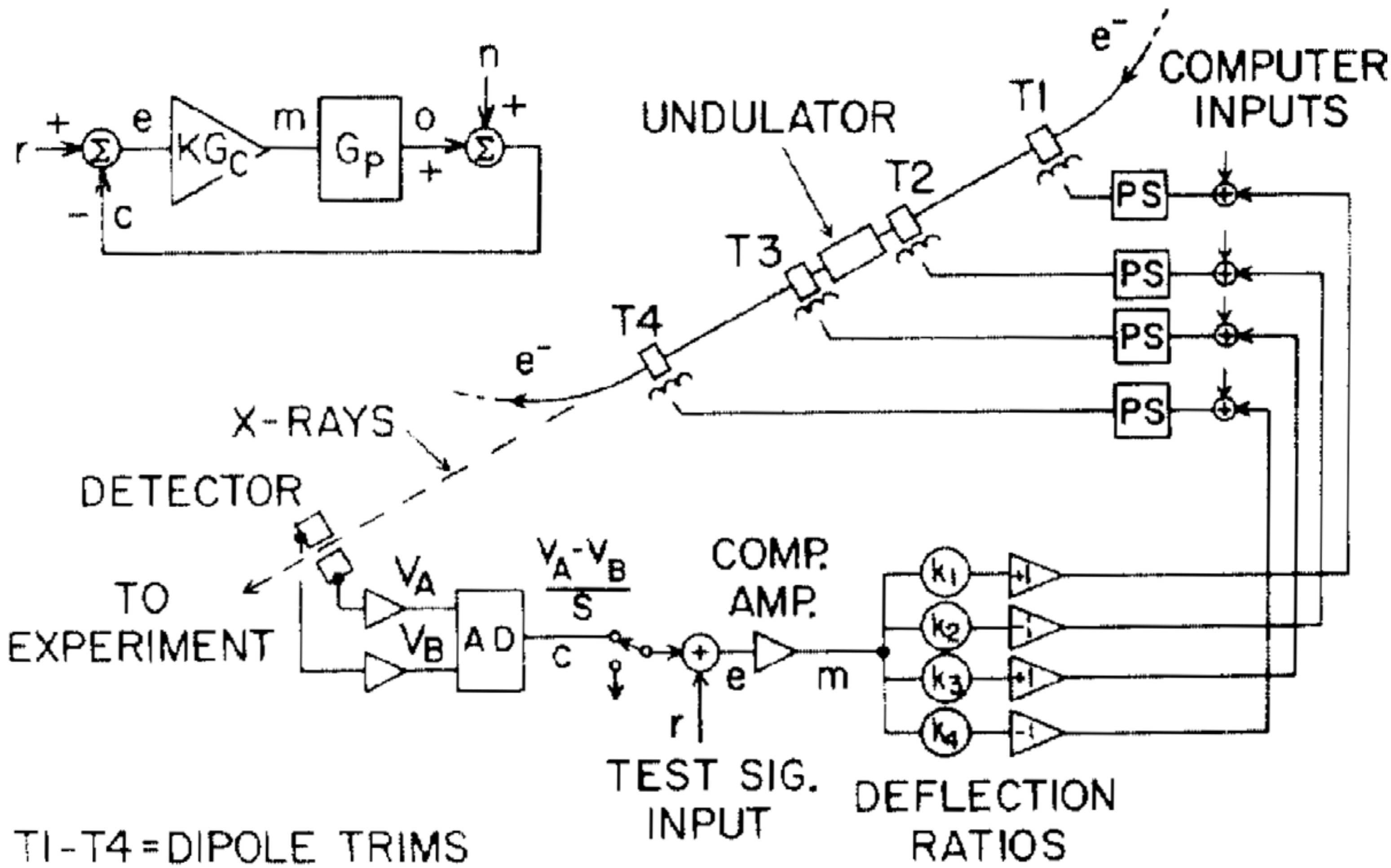
Local orbit correction using a four-corrector local bump

- The three-corrector bump does not have sufficient degrees of freedom to control both position and angle of the electron orbit. Superimposing two three-corrector local bumps to create a four-corrector bump provides independent control of both position and angle



- Bump coefficients for the four-corrector bump can be calculated via superposition and the three-corrector formulae.

NSLS Local Correction System



Perfect local control would be ideal

- Local control allows exact control of electron orbit through the x-ray source point. So why do anything else?

Problem #1

- Local bumps provide exact control of beam position and angle.
- Can use rf bpms (electron orbit) or photon bpms (the ideal).
- But, it is very difficult to exactly measure the electron or photon beam position with the required long-term and short-term performance.
- Without inadequate measurements of beam position, local control does an excellent job of point in the wrong direction.

Problem #2

- Requires many correctors (4 per x-ray source point) and many bpms (two per x-ray source point).
- Cost factor
- Space factor in lattice
- Fighting between loops, because it is impossible to completely decouple the local loops from the global orbit.

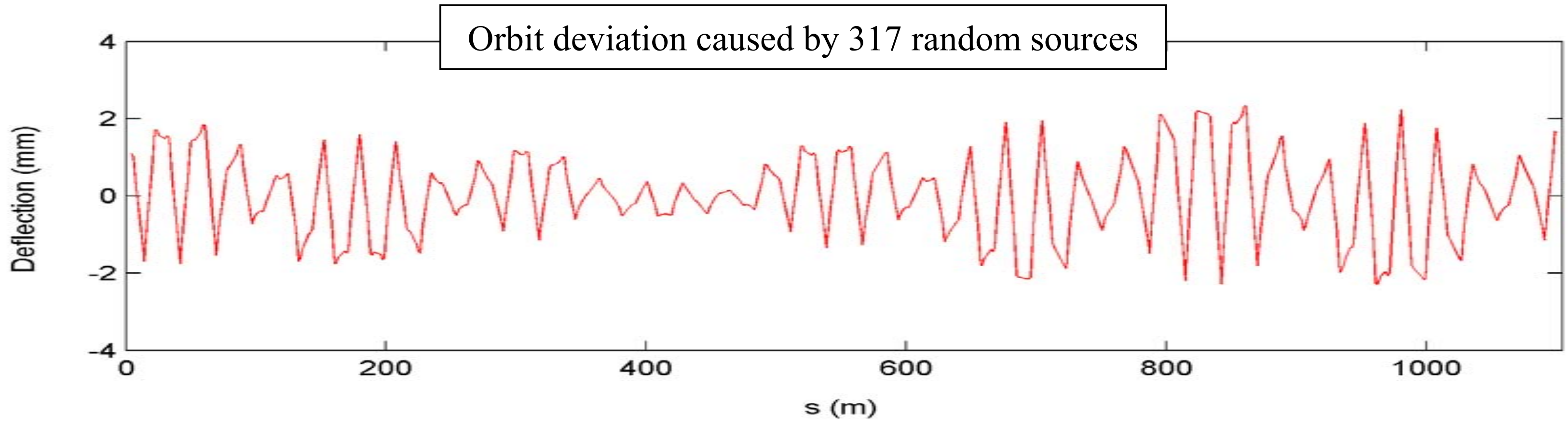
Global orbit correction

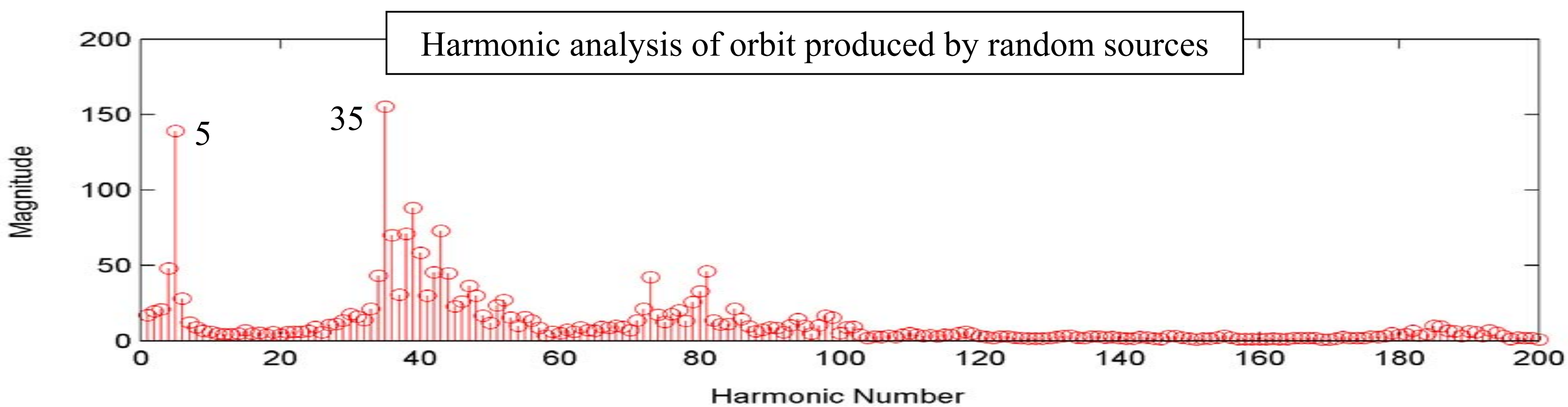
- Global correction algorithms take advantage of the global nature of disturbances from individual sources to correct everywhere.
- The goal is to find a set of corrector values that best reproduce the measured orbit deviations, and use that set for cancellation.

- **Harmonic correction** – utilize harmonic nature of orbit disturbances.
- **Response matrix inversion** – relates bpm and corrector response vectors.

Harmonic orbit correction

- Spatial Fourier analysis of orbit motion from many random disturbances shows that harmonics around the betatron tune dominate.
- Orbit correction can be performed by globally canceling those harmonics.





NSLS VUV-ring analog harmonic orbit feedback system

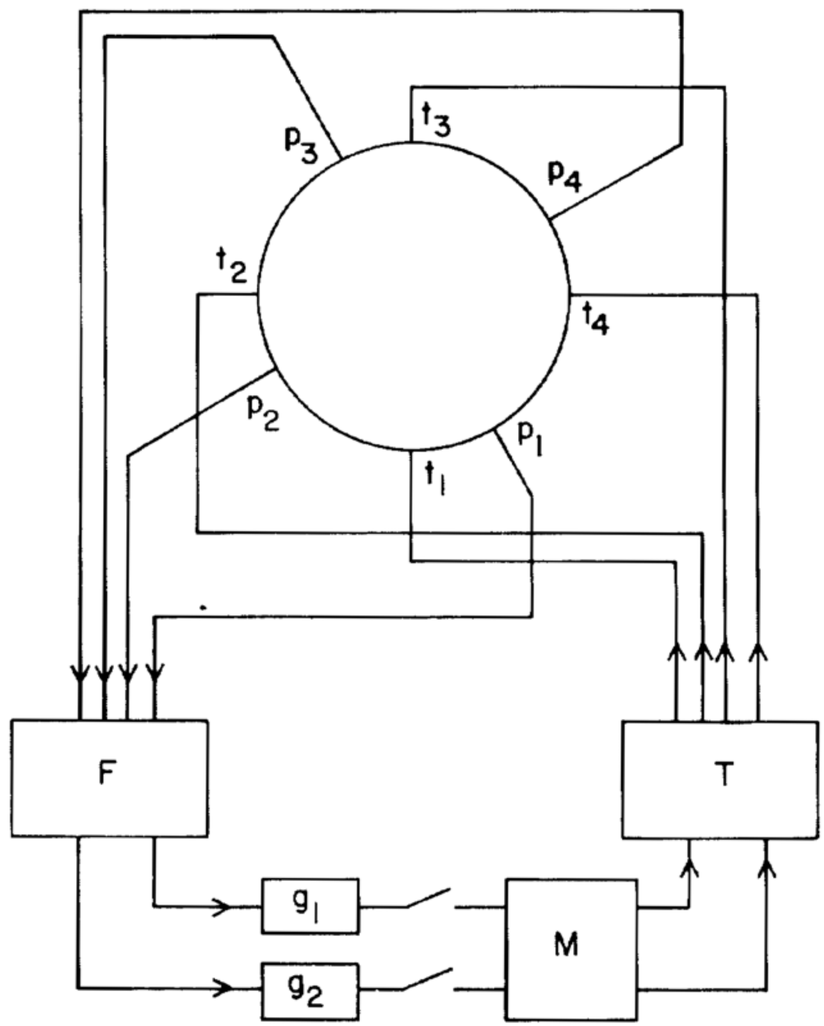
- For betatron tune of 1.2, orbit distortion is approximated to:

$$\eta \approx a \sin\varphi + b \cos\varphi$$

- Measure strength of ***a*** & ***b*** coefficients from least-squares fit to four bpm signals.
- Drive four correctors based on measured strength of ***a*** & ***b*** coefficients.

$$\begin{bmatrix} a \\ b \end{bmatrix} = F \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} \qquad \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{bmatrix} = T \begin{bmatrix} -a \\ -b \end{bmatrix}$$

- ***F*** and ***T*** are computed from Courant-Snyder phases at the pickups and trims.
- ***M*** diagonalizes the control loops.
- ***F***, ***T***, ***M*** are implemented as analog networks.



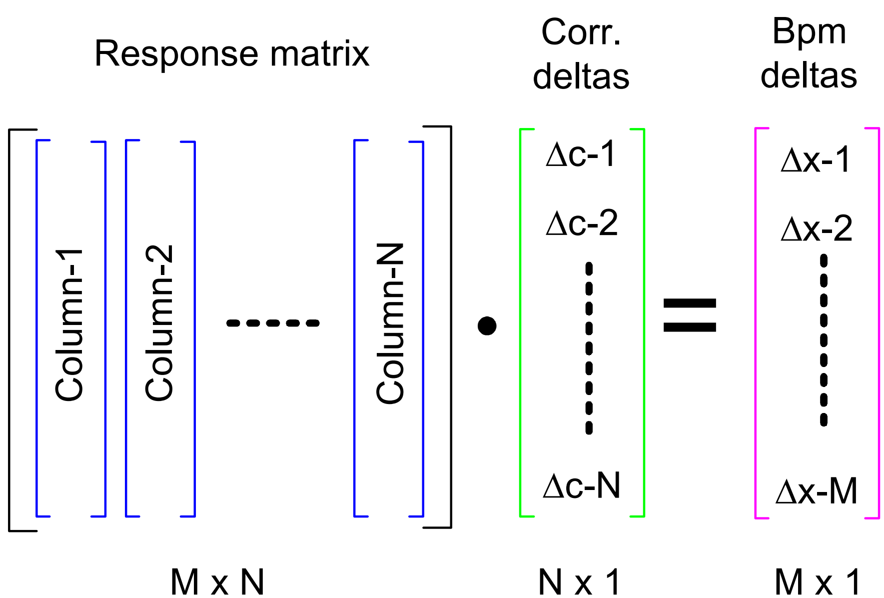
LH Yu - 1989

Response Matrix Equation

- The response matrix equation describes how the orbit changes at specific bpm locations for small changes in strength of specific correctors

$$R \cdot \Delta c = \Delta x$$

- Each column of the response matrix defines the change in bpm readings for a unit change in a given corrector



Obtaining the response matrix

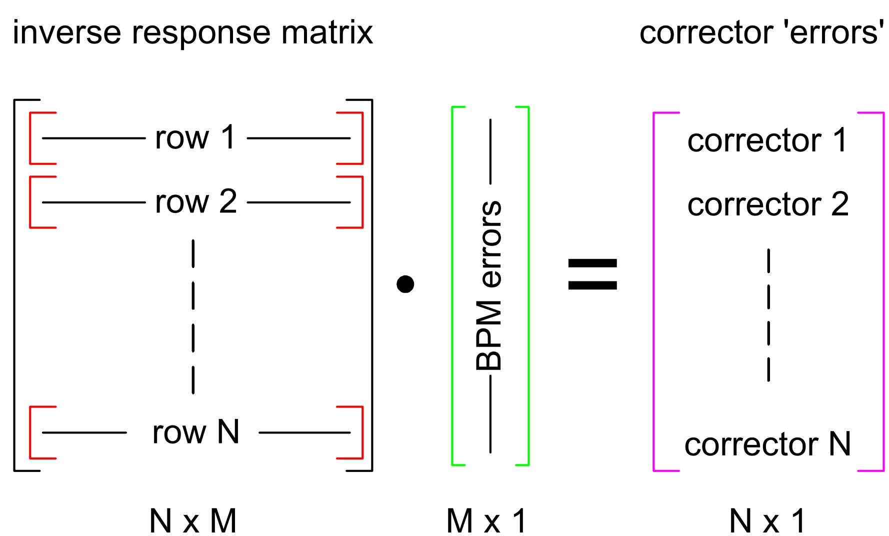
- **Measure it…**
  - Change the strength of one corrector at a time, and measure the resulting orbit change at the bpms of interest.
  - Reflects how the system actually behaves.
  - Subject to measurement errors and systematic effects.
  - Time-consuming to get good data.
  - Accuracy and speed of measurement can be significantly improved by doing the measurement with AC signals (APS RT-Feedback “AC lock-in”

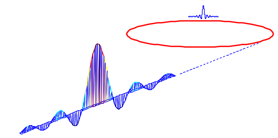
- **From a model of the accelerator lattice…**
  - Describes how the system should behave, not how it actually behaves.
  - Not subject to measurement errors or systematics.
  - Models are usually refined by including actual data from magnet measurements and survey/alignment, baseline orbits, and many other measured parameters. Results can be extremely accurate.

Inverse response matrix equation

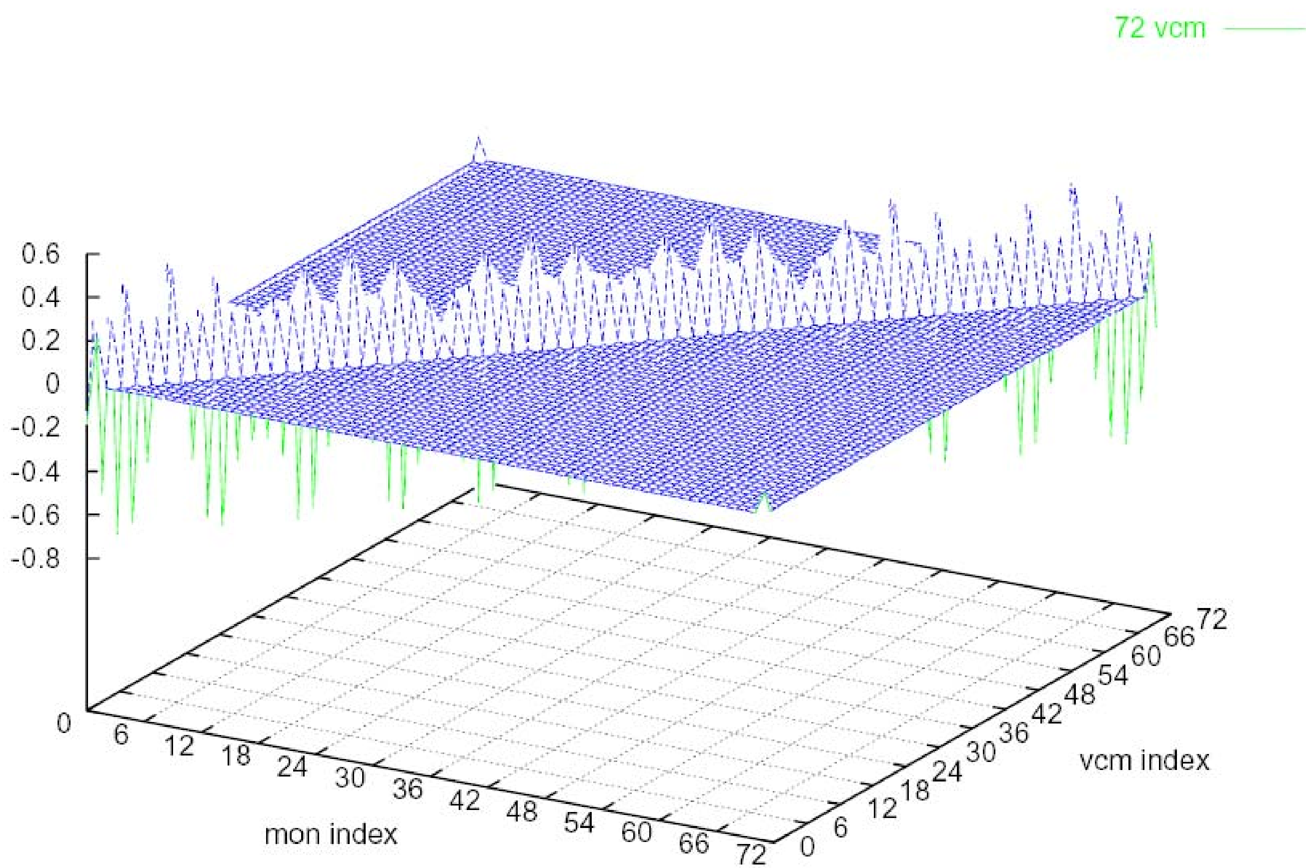
- To correct the orbit, we need the ‘inverse response matrix,’ which maps the measured orbit deviations at specific bpms to the changes in corrector strength needed to correct them

$$R^{-1} \cdot \Delta x = \Delta c$$





Diagonal nature of inverse response matrix (SLS)



Inverting the response matrix

- The response matrix equation describes a set of simultaneous linear equations. The inverse response matrix is the solution to that set of equations.

Most-commonly used techniques at light-sources

- Singular-value decomposition (SVD).
- Least-squares pseudo-inverse.
- Weighted least-squares pseudo-inverse.

Factors that influence the approach to inverting the response matrix

- Relative number of correctors and bpms
  - Same number of orthogonal correctors as bpms = exact solution possible.
  - More correctors than bpms = over-determined solution (many solutions).
  - More bpms than correctors = under-determined solution (no exact solution).
- Correction goals (eg exact control of photon source-points, rms correction).

Three cases of the response matrix dimensions

- Case 1 – same number of bpms as correctors
  - Matrix is square, and potentially invertible.
  - If the matrix is invertible, then it generates an exact solution at all bpms.

- Case 2 – more bpms that correctors
  - Matrix is rectangular and not invertible.
  - More equations that variables, so there is no exact solution.
  - Can generate least-squares solution to bpm readbacks:
    - Pseudo-inverse matrix, provided the matrix has independent columns.
    - SVD, with or without removal of singular values.
  - Solution will be more robust against individual bpm errors because of least-squares fit.

- Case 3 – more correctors than bpms
  - Matrix is rectangular and not invertible.
  - More variables than equations, so there are many exact solutions.
  - Can generate least-squares solution to corrector drives:
    - Pseudo-inverse matrix, provided the matrix has independent columns.
    - SVD, with or without removal of singular values.
  - Generates exact solution at all bpm locations, with rms corrector power minimized, and with rms orbit between bpms lower than for case #1 because of additional correctors.

Why the exact solution may not be what we want…

- With the same number of suitably chosen correctors and bpms, the response matrix inverse can allow exact correction at all bpms simultaneously.

This may not be such a good idea…

- Creating an exact solution to the response matrix equation constrains the orbit such that the error at each bpm is exactly zero, but orbit between the bpms is unconstrained.
- Orbit measurement errors, or differences between the response matrix model and the physical lattice (that can occur for many reasons), can result in an orbit that deviates substantially from the desired orbit between the bpms.

- Any measurement errors or measurement noise will be directly translated into positioning errors of the particle beam at the bpm locations.

Selection of bpms and correctors

- Effective orbit correction relies on being able to decompose the measured orbit into spatial modes that are strongly coupled to available correctors or combinations of correctors.

- Modal decomposition options
  - Spatial Fourier transform (harmonic decomposition)
  - Eigenvector spectral decomposition
  - SVD spectral decomposition

Singular Value Decomposition

- Applicable to both square and non-square matrices.
- Offers control over the robustness of the matrix inverse (condition number).

- Any *MxN* matrix, ***R*** can be decomposed into a product of three matrices:

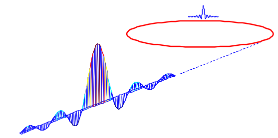
$$R = U \cdot S \cdot V^T$$

Where ***U*** and ***V*** are unitary matrices, and ***S*** is a diagonal matrix.

- The diagonal elements of ***S*** are the square roots of all the non-zero eigenvalues of both $\boldsymbol{RR^T}$ and $\boldsymbol{R^TR}$
- The columns of ***U*** are the normalized eigenvectors of $\boldsymbol{RR^T}$
- The columns of ***V*** are the normalized eigenvectors of $\boldsymbol{R^TR}$

e.g.,

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{21} \\ u_{21} & u_{22} \end{bmatrix} \cdot \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \end{bmatrix} \cdot \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{bmatrix}^T$$



Matrix inversion using SVD

- SVD decomposition…

$$R = U \cdot S \cdot V^T$$

- Since a unitary matrix comprises orthogonal columns, all of unit length, then

$$U \cdot U^{-1} = U \cdot U^T = U^T \cdot U = 1$$

- The matrix inversion is therefore

$$R^{-1} = \left(U \cdot S \cdot V^T\right)^{-1} = V \cdot S^{-1} \cdot U^T$$

SVD as a vector-space transformation

- Using the SVD formulation, the standard response matrix equation

$$R \cdot \Delta c = \Delta x$$

becomes

$$U \cdot S \cdot V^T \cdot \Delta c = \Delta x$$

or

$$S \cdot V^T \cdot \Delta c = U^T \cdot \Delta x$$

SVD as a vector-space transformation (cont)

- This can be viewed as a vector transformation of the corrector and bpm spaces

$$\begin{bmatrix} s_{11} & 0 & 0 \\ 0 & s_{22} & 0 \\ 0 & 0 & s_{33} \end{bmatrix} \cdot \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ v_{31} & v_{32} & v_{33} \end{bmatrix}^T \cdot \begin{bmatrix} \Delta c_1 \\ \Delta c_2 \\ \Delta c_3 \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{bmatrix}^T \cdot \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \end{bmatrix}$$

Or…

$$\begin{bmatrix} s_{11} & 0 & 0 \\ 0 & s_{22} & 0 \\ 0 & 0 & s_{33} \end{bmatrix} \cdot \begin{bmatrix} \Delta tc_1 \\ \Delta tc_2 \\ \Delta tc_3 \end{bmatrix} = \begin{bmatrix} \Delta tx_1 \\ \Delta tx_2 \\ \Delta tx_3 \end{bmatrix}$$

- Since the coupling matrix ***S*** is diagonal, the transformed vector spaces are decoupled (ie orthogonal).

Using SVD to improve the robustness of the matrix inverse

- The 'S' matrix in the SVD decomposition is a diagonal matrix, with each element representing the strength of the coupling between transformed corrector and bpm vectors.

$$\begin{bmatrix} s_{11} & 0 & 0 \\ 0 & s_{22} & 0 \\ 0 & 0 & s_{33} \end{bmatrix} \cdot \begin{bmatrix} \Delta tc_1 \\ \Delta tc_2 \\ \Delta tc_3 \end{bmatrix} = \begin{bmatrix} \Delta tx_1 \\ \Delta tx_2 \\ \Delta tx_3 \end{bmatrix}$$

- The inverse relationship is obtained simply by inverting the diagonal matrix

$$\begin{bmatrix} \Delta tc_1 \\ \Delta tc_2 \\ \Delta tc_3 \end{bmatrix} = \begin{bmatrix} \Delta tx_1 \\ \Delta tx_2 \\ \Delta tc_3 \end{bmatrix} \cdot \begin{bmatrix} 1/s_{11} & 0 & 0 \\ 0 & 1/s_{22} & 0 \\ 0 & 0 & 1/s_{33} \end{bmatrix}$$

But if one of the diagonal elements of ***S*** is close to zero, the inverse matrix will be singular, and the orbit correction algorithm will be unstable.

Removing unstable (singular) modes from the inverse

- Before performing the matrix inversion, zero out small diagonal elements of the diagonal matrix, and cancel associated rows of V,

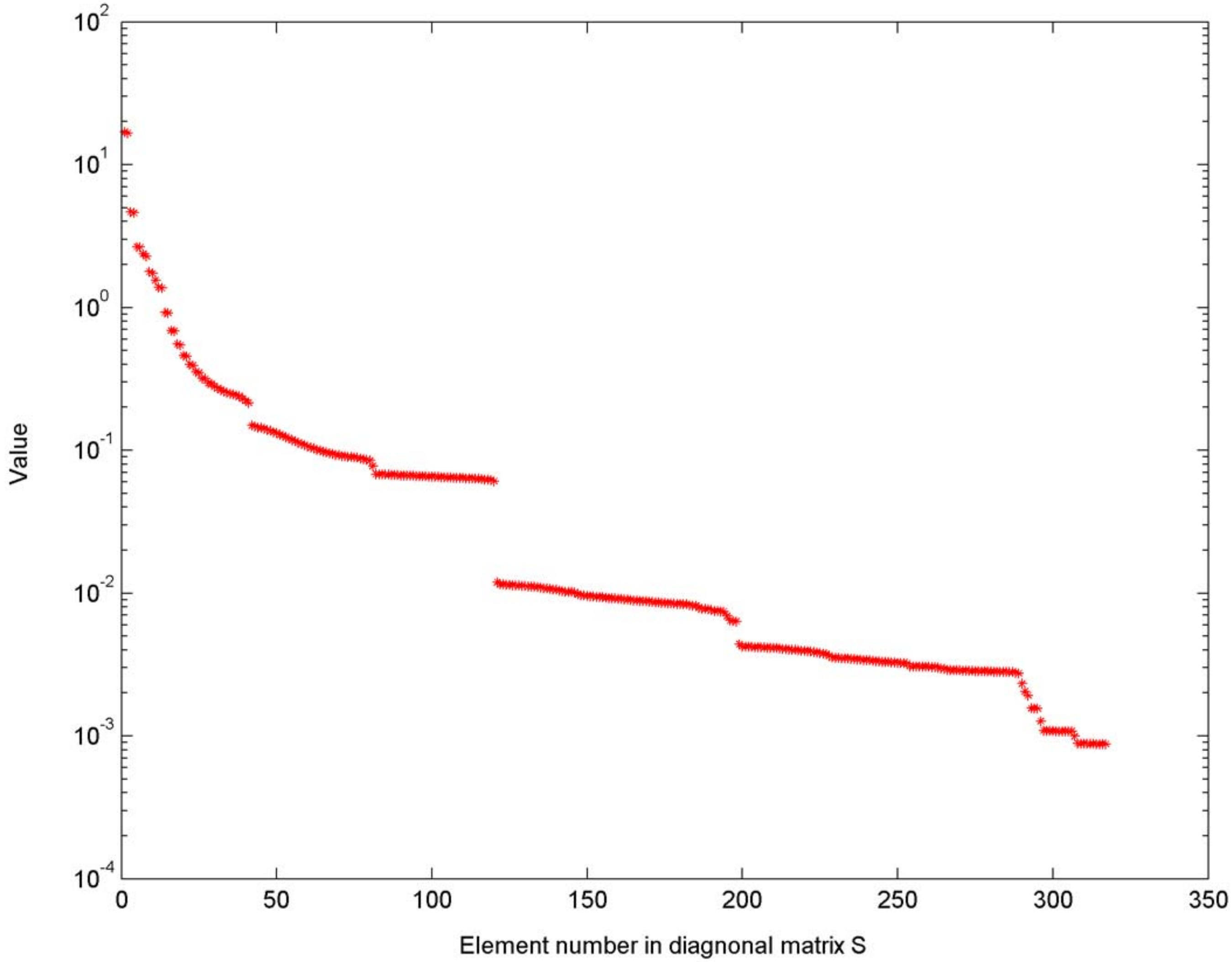
e.g.,

$$\begin{bmatrix} s_{11} & 0 & 0 \\ 0 & s_{22} & 0 \\ 0 & 0 & \overset{0}{s_{33}} \end{bmatrix} \cdot \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \\ \cancel{v_{31}} & \cancel{v_{32}} & \cancel{v_{33}} \end{bmatrix}^T \cdot \begin{bmatrix} \Delta c_1 \\ \Delta c_2 \\ \Delta c_3 \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{bmatrix}^T \cdot \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta c_3 \end{bmatrix}$$

Resulting in…

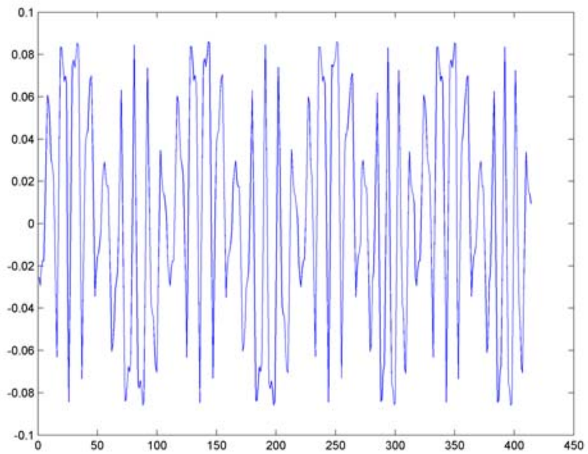
$$\begin{bmatrix} s_{11} & 0 \\ 0 & s_{22} \end{bmatrix} \cdot \begin{bmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \end{bmatrix}^T \cdot \begin{bmatrix} \Delta c_1 \\ \Delta c_2 \\ \Delta c_3 \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{bmatrix}^T \cdot \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta c_3 \end{bmatrix}$$

Singular values for APS horizontal response matrix (414 bpms x 317 correctors)

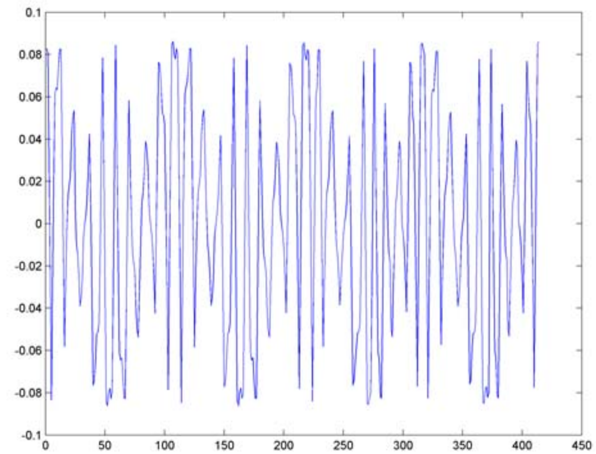


Some BPM modes for APS horizontal response matrix (414 bpms x 320 correctors)

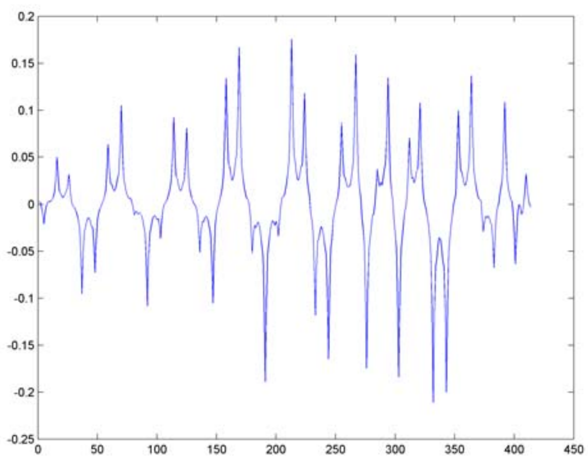
Col #1 from ***U*** matrix



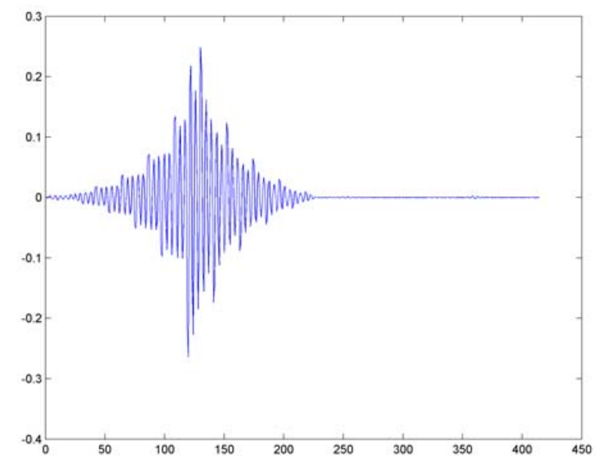
Col #2 from ***U*** matrix



Col #100 from ***U*** matrix



Col #200 from ***U*** matrix



Applications

Number of bpms = number of correctors

- With the right selection of correctors, this offers possibly of exactly correcting each bpm location. For reasons discussed shortly, this may not be desirable.
- Remove small singular values to convert exact solution to a least-squares solution, while reducing required rms corrector power.

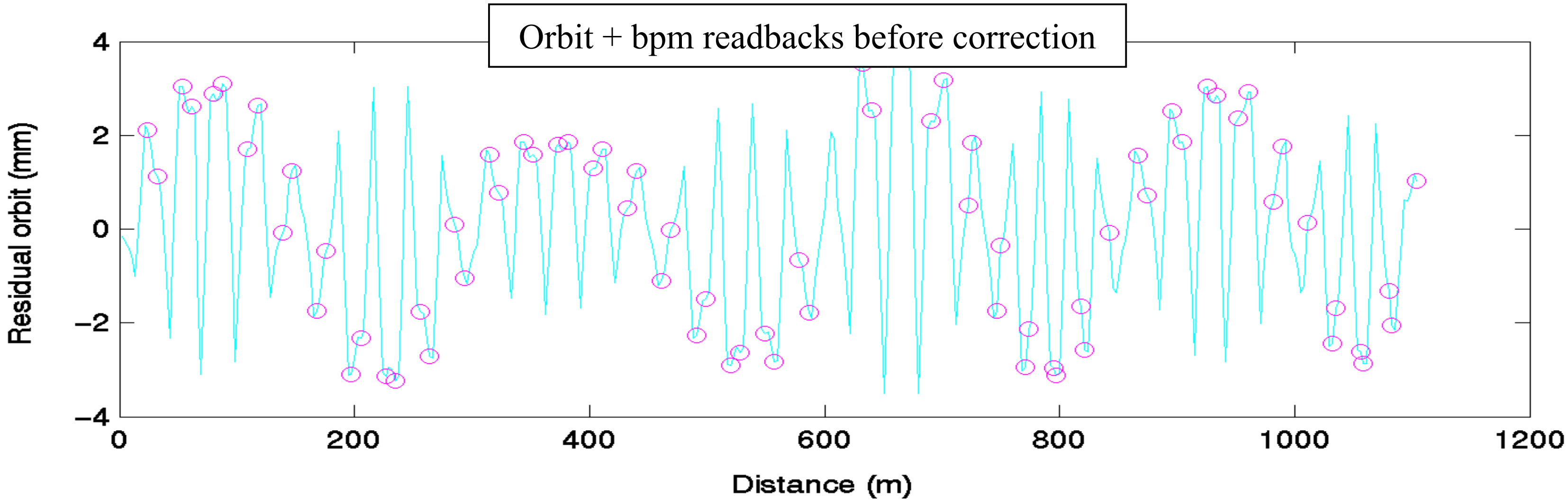
Number of bpms > number of correctors

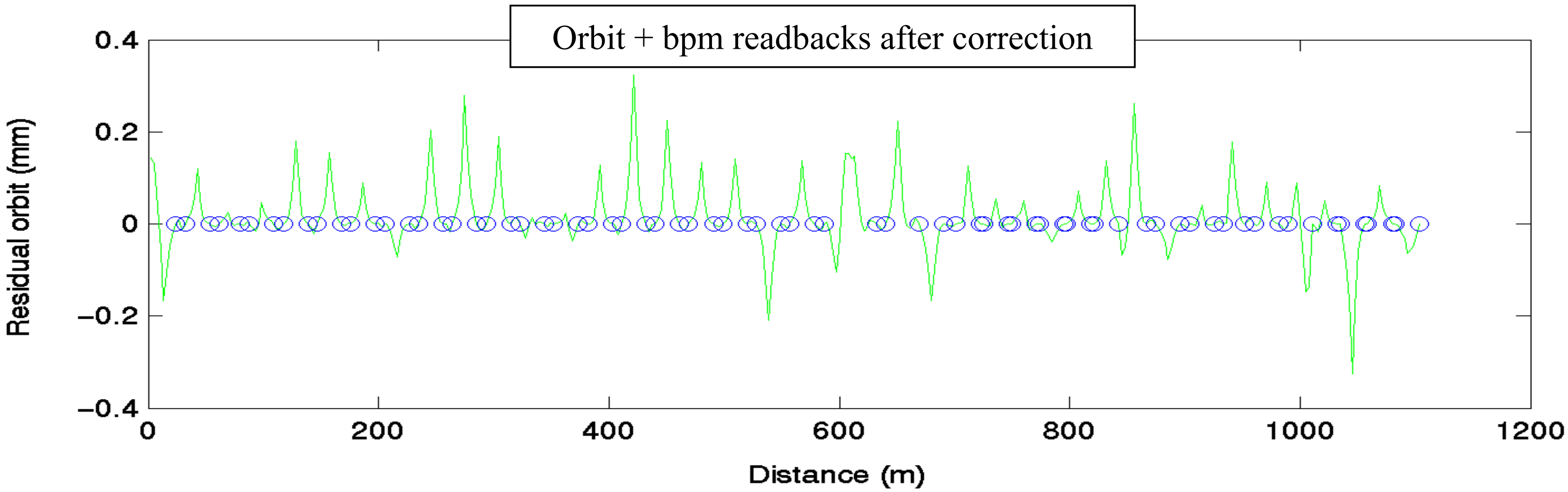
- No exact solution (insufficient degrees of freedom).
- Leaving all singular values results in a least-squares solution.
- Robustness can be improved further by removing the smaller singular values.

Number of bpms < number of correctors

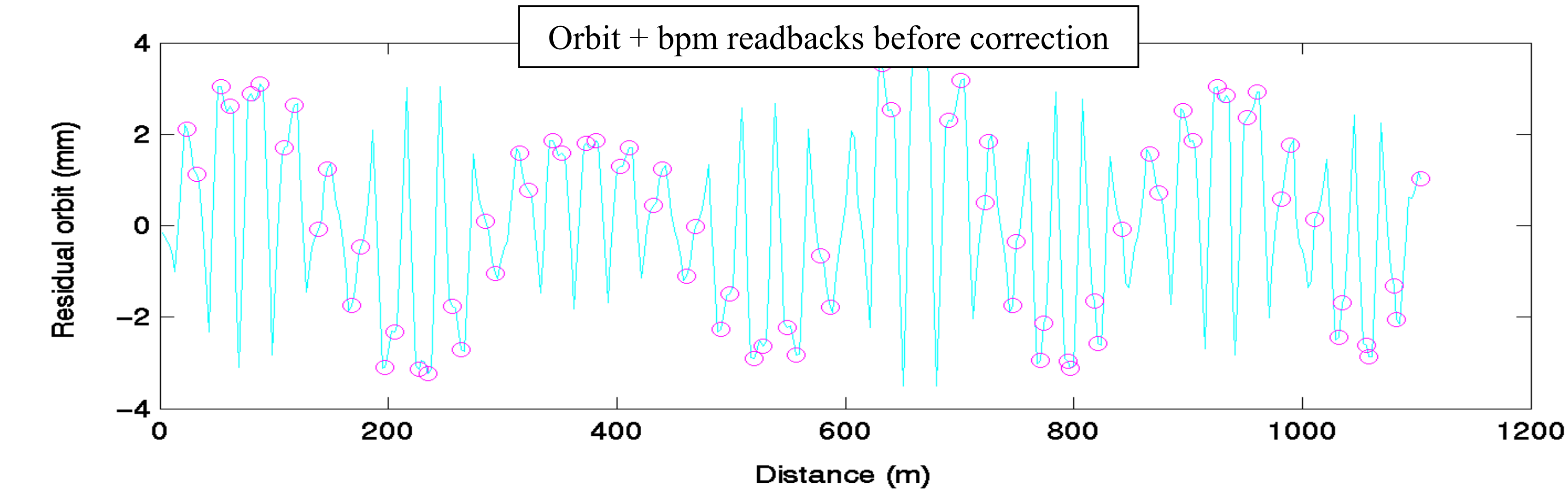
- Many exact solutions.
- Leaving all singular values results in rms minimization of corrector power.
- Robustness can be improved further by removing the smaller singular values.

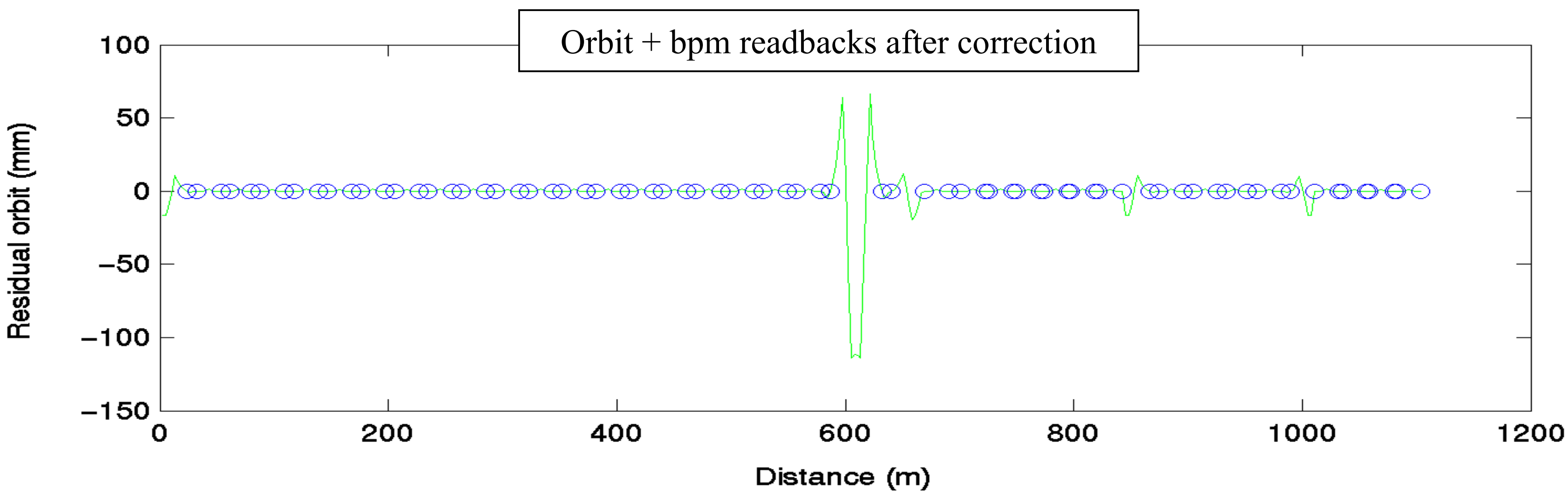
Robust set of 78 bpms + 79 correctors (all SVs)



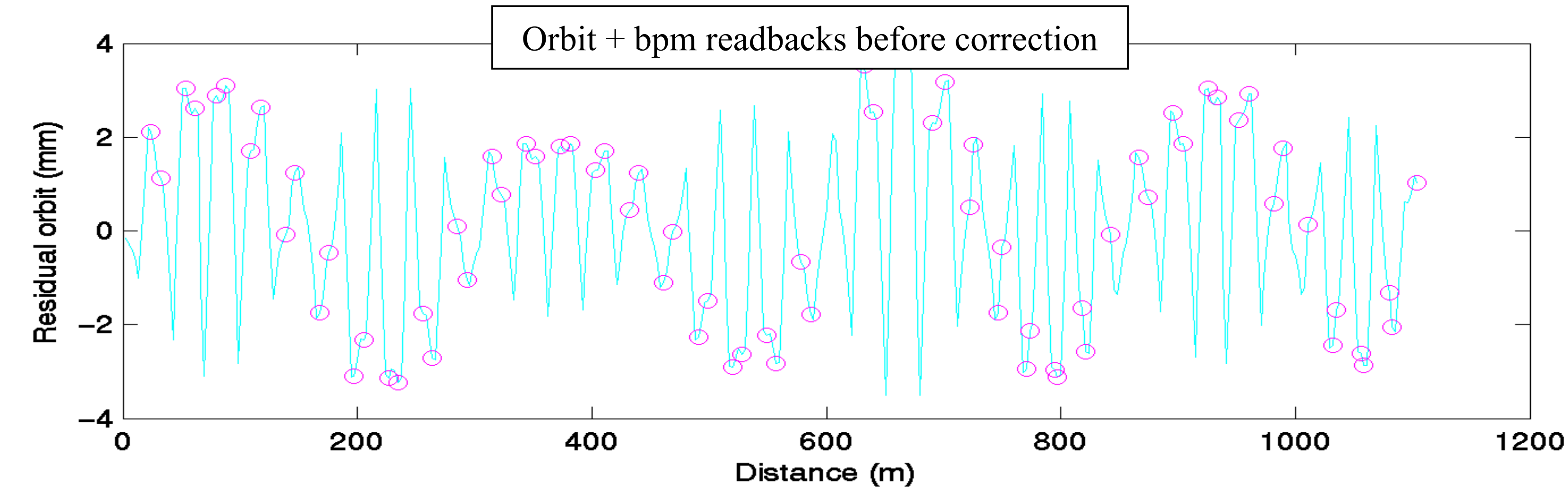


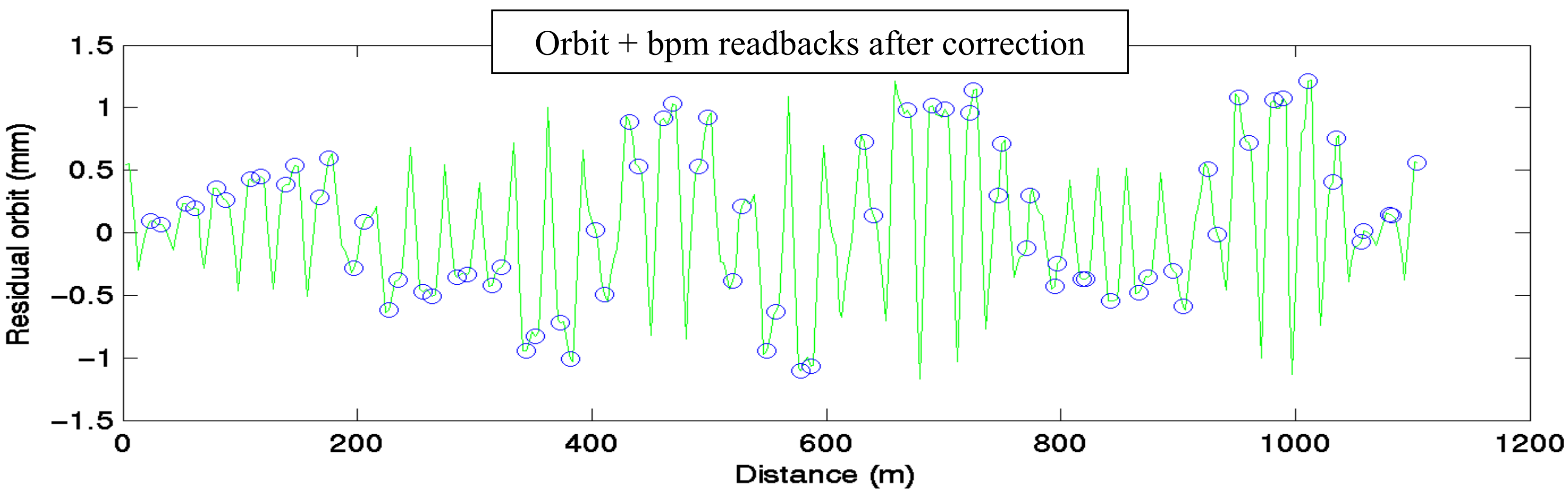
Singular set of 78 bpms + 79 correctors (all SVs)



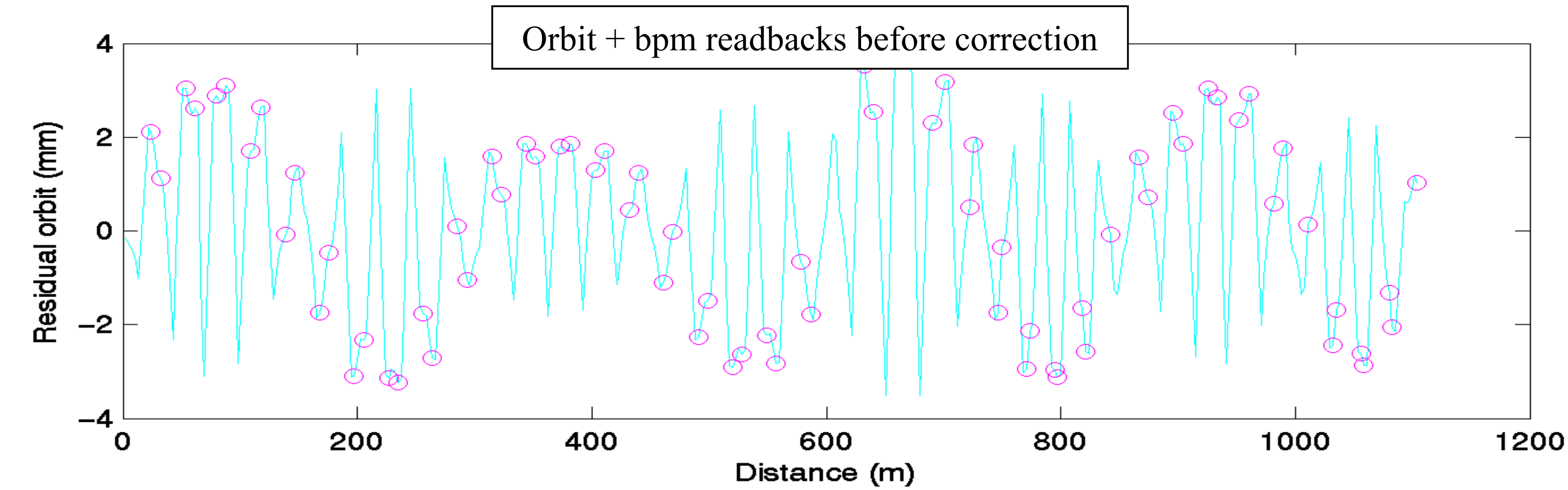


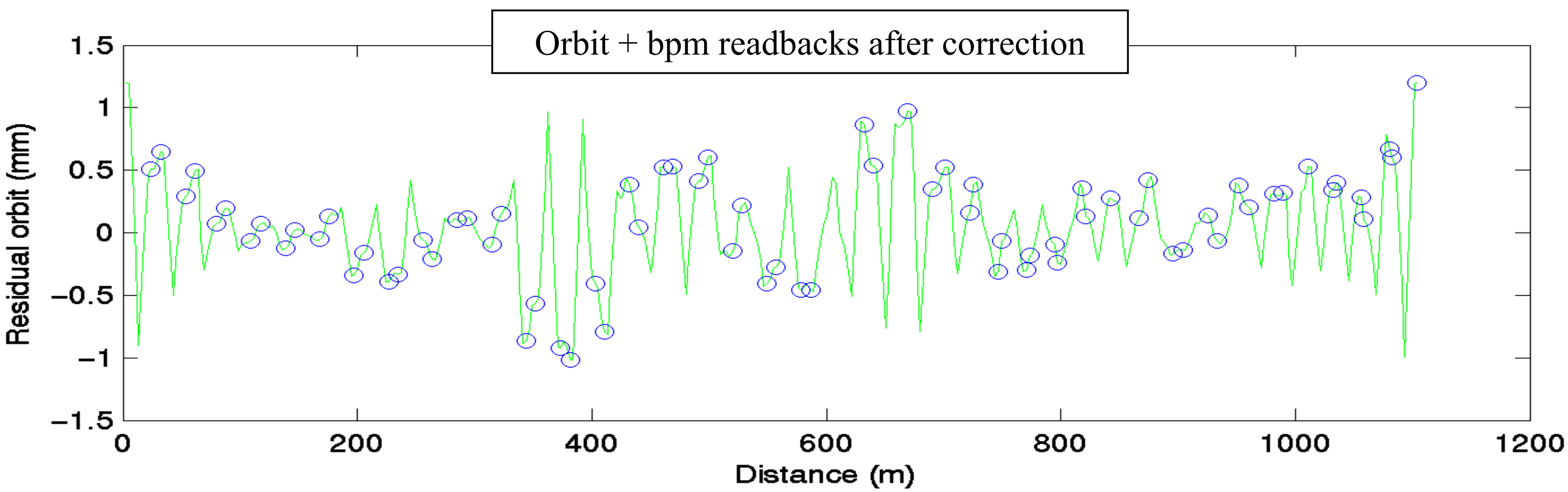
Singular set of 78 bpms + 79 correctors (1 SV)



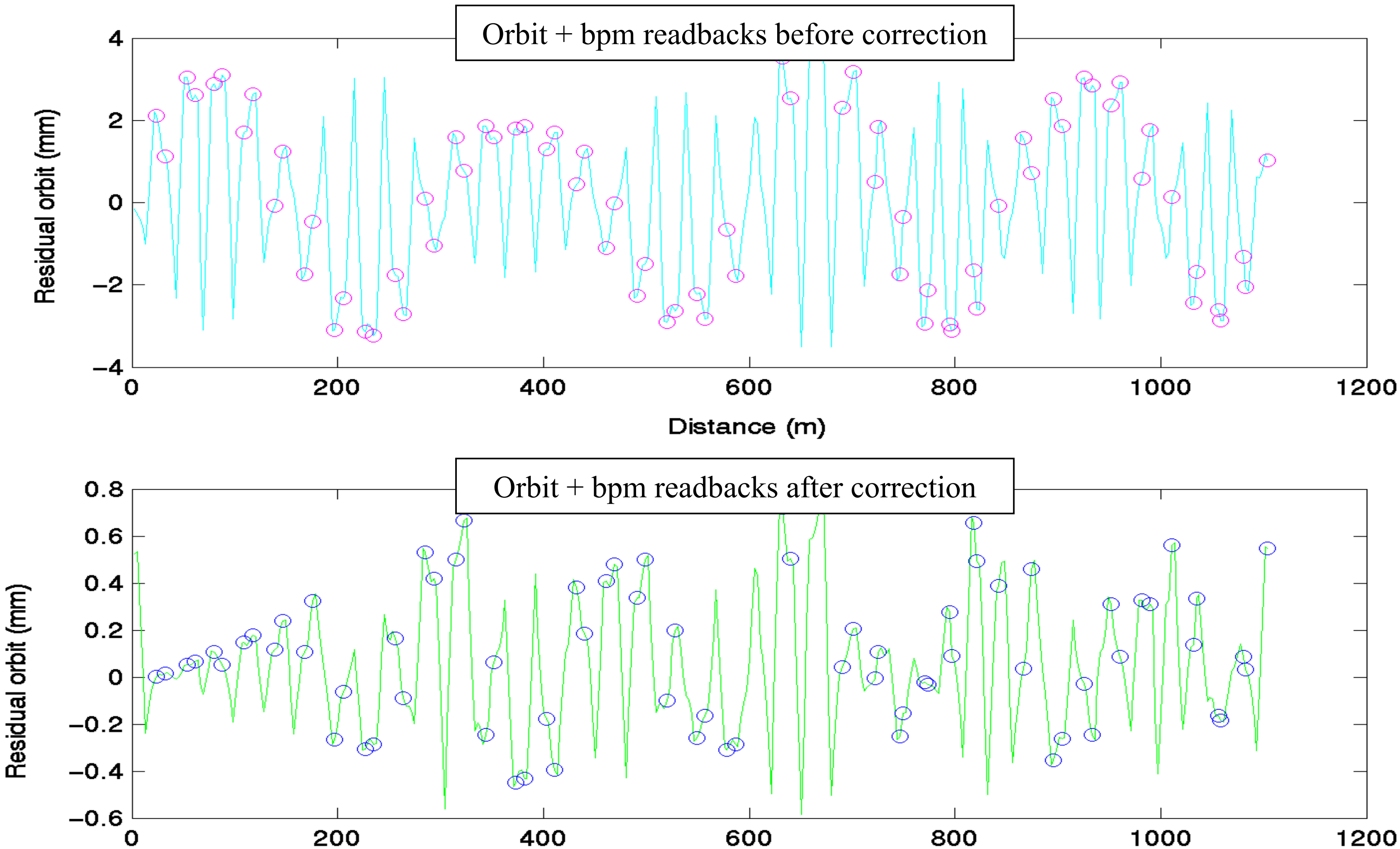


Singular set of 78 bpms + 79 correctors (5 SVs)

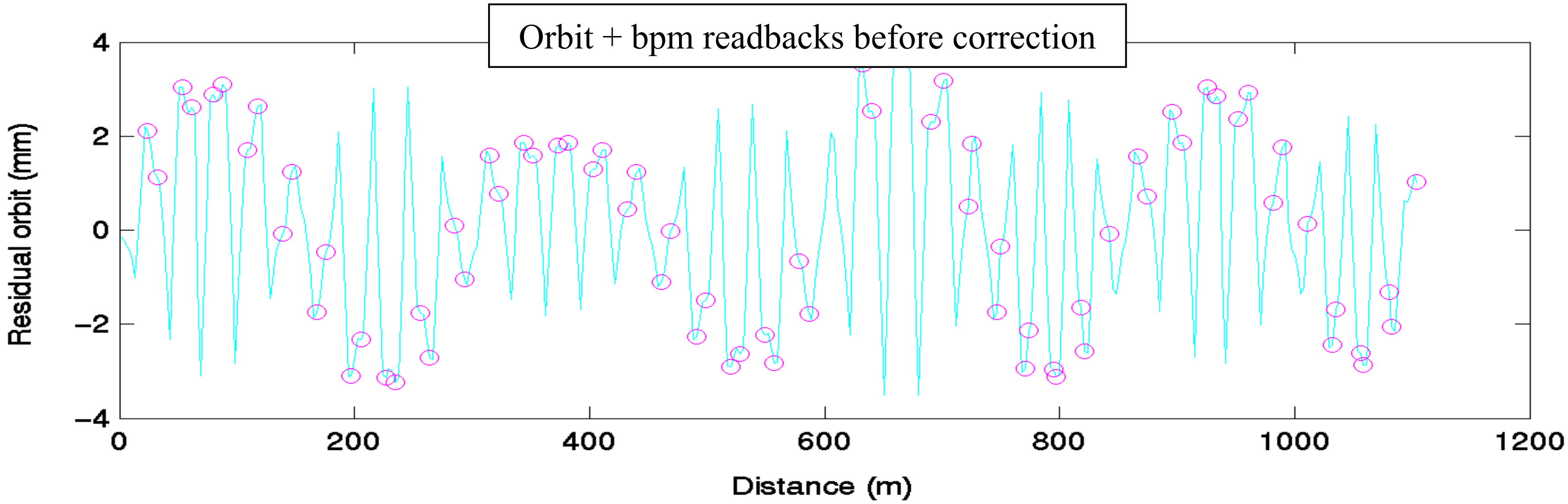


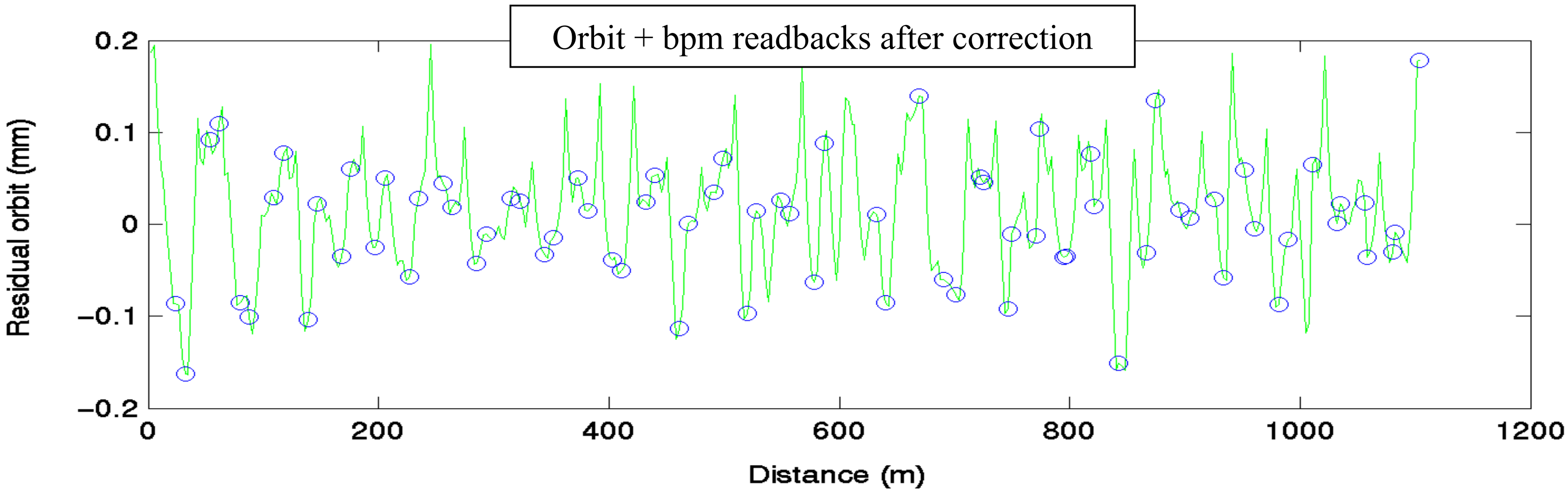


Singular set of 78 bpms + 79 correctors (10 SVs)

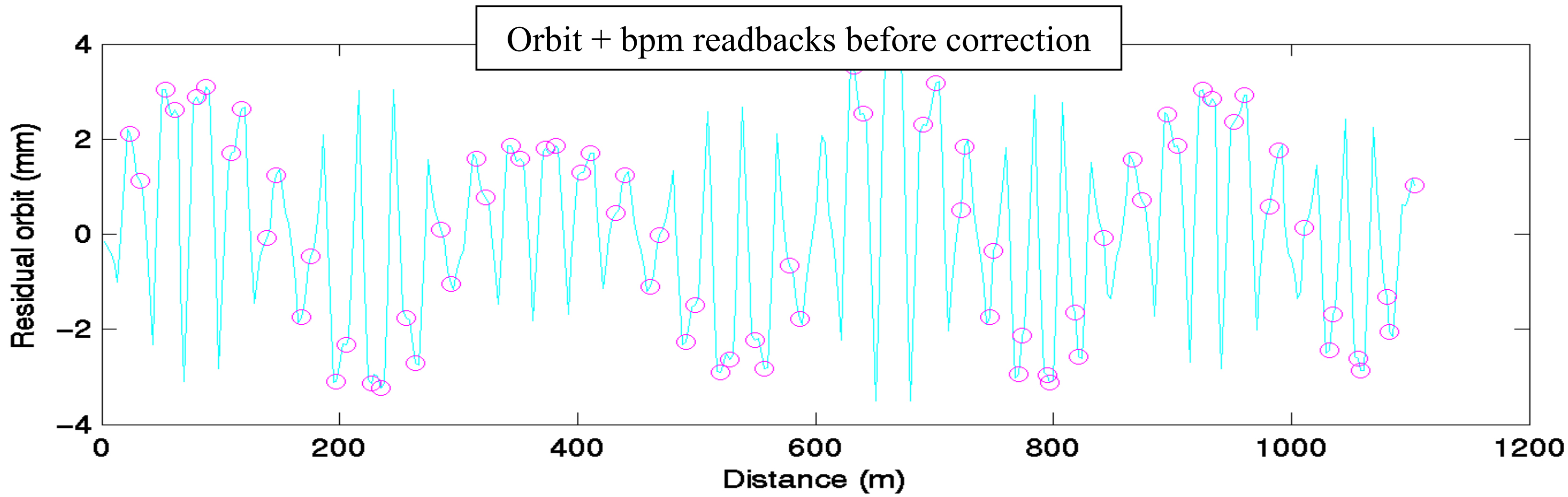


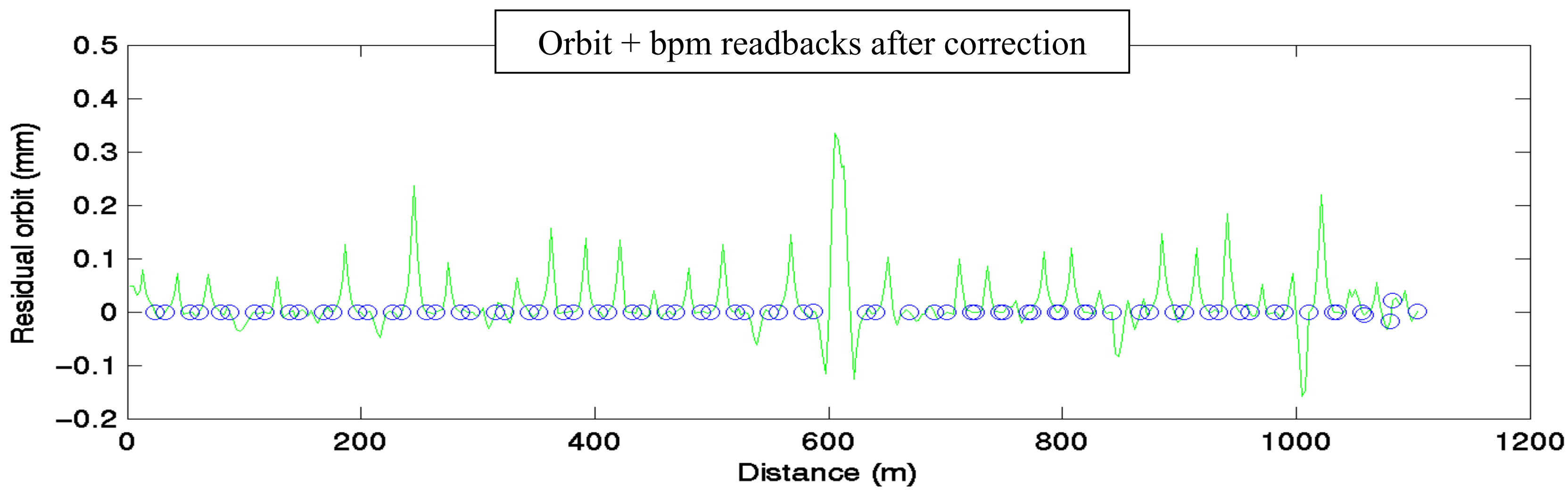
Singular set of 78 bpms + 79 correctors (30 SVs)





Singular set of 78 bpms + 79 correctors (73 SVs)

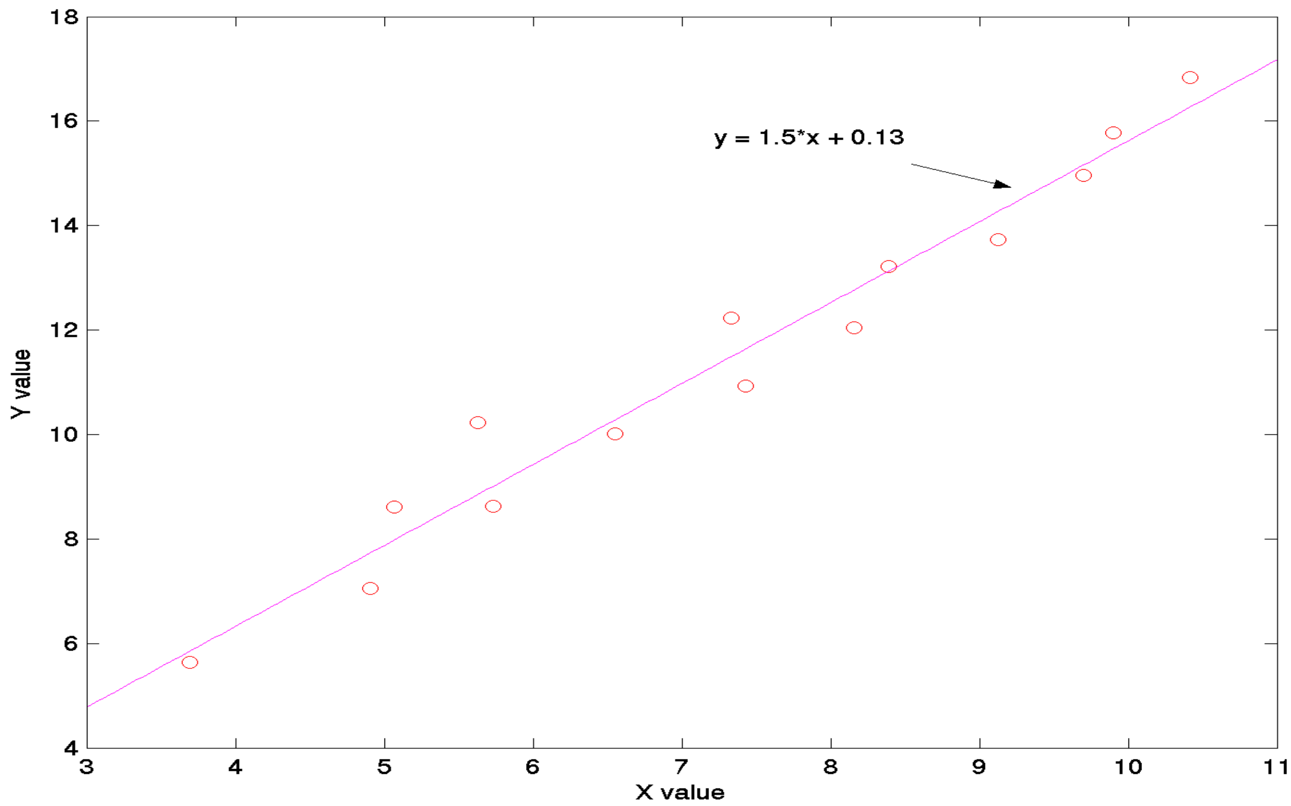




Least-squares matrix inversion

- Instead of removing unstable singular values, the robustness of the orbit correction algorithm is significantly improved when there are many more bpms than correctors used in the orbit correction algorithm.
  - APS has over 400 available rf bpms, and typically uses only 80 correctors for orbit correction (per plane)
  - This is not possible at all light-sources due to physical space in the lattice or to budgetary constraints.
- In linear algebra terms, having more bpms than correctors results in a under-determined (over-constrained) system of equations (there are more measurements than unknowns), for which there is no exact solution.

Least-squares application to curve-fitting



Least-squares matrix inversion

Method 1

- Use the SVD formulation, retaining all singular values in the inversion process.

Method 2

- Pseudo-inverse formulation

$$R \cdot \Delta c = \Delta x$$

$$\therefore R^T \cdot R \cdot \Delta c = R^T \cdot \Delta x$$

$$\therefore \Delta c = \left(R^T \cdot R \cdot\right)^{-1} R^T \cdot \Delta x$$

$$\therefore R_{pinv} = \left(R^T \cdot R \cdot\right)^{-1} R^T$$

Conditions when the pseudo-inverse fails

- Mathematically, the pseudo-inverse formulation fails when columns or rows of the response matrix are not independent.

- Physical connection with non-independence of response matrix rows
  - Columns of the response matrix are the individual bpm signatures for each of the chosen correctors.
  - Columns not being independent implies that the bpm responses from different correctors are the same within measurement uncertainties.
  - This is possible for small matrices, or for bpm sets that are poorly chosen or poorly located within the accelerator lattice.
  - It could also happen for poorly chosen and/or poorly located correctors.

- The SVD formulation still allows a method for inverting such poorly selected bpms

Weighted least-squares

- The weighted least-squares matrix inversion formulation allows the orbit to be corrected exactly at specific locations, eg at photon source points, or preferential weighting of specific bpms, eg because they have lower noise floors.
- By suitably weighting either photon bpms or rf bpms around the photon source-points, the weighted least-squares allows integration of local exact control with global rms control, using a single correction algorithm.

- Given 'good' readbacks of photon and/or electron beam position at the photon beamlines, this is arguably the best algorithm to use.
- It eliminates difficulties with running independent global and local correction algorithms that can (and do) fight each other.

- Weighted least-squares algorithm is now in use at the APS, where bending magnet photon bpms are integrated into the global orbit correction algorithm, and are given heavy weights, so as to correct exactly at those locations, while simultaneously correcting the electron orbit in a global rms sense.

Weighted least-squares

- Simply use the standard algorithm from linear algebra.

- Starting with the response matrix equation, add a diagonal matrix ***W*** containing weighting factors for each of the bpm readings

$$R \cdot \Delta c = \Delta x$$

$$W \cdot R \cdot \Delta c = W \cdot \Delta x$$

- Then invert using the least-squares formulation…

$$\therefore (W \cdot R \cdot)^T \cdot W \cdot R \cdot \Delta c = (W \cdot R \cdot)^T \cdot W \cdot \Delta x$$

$$\therefore R_{winv} = \left((W \cdot R \cdot)^T \cdot W \cdot R \cdot\right)^T \cdot (W \cdot R \cdot)^T \cdot W$$

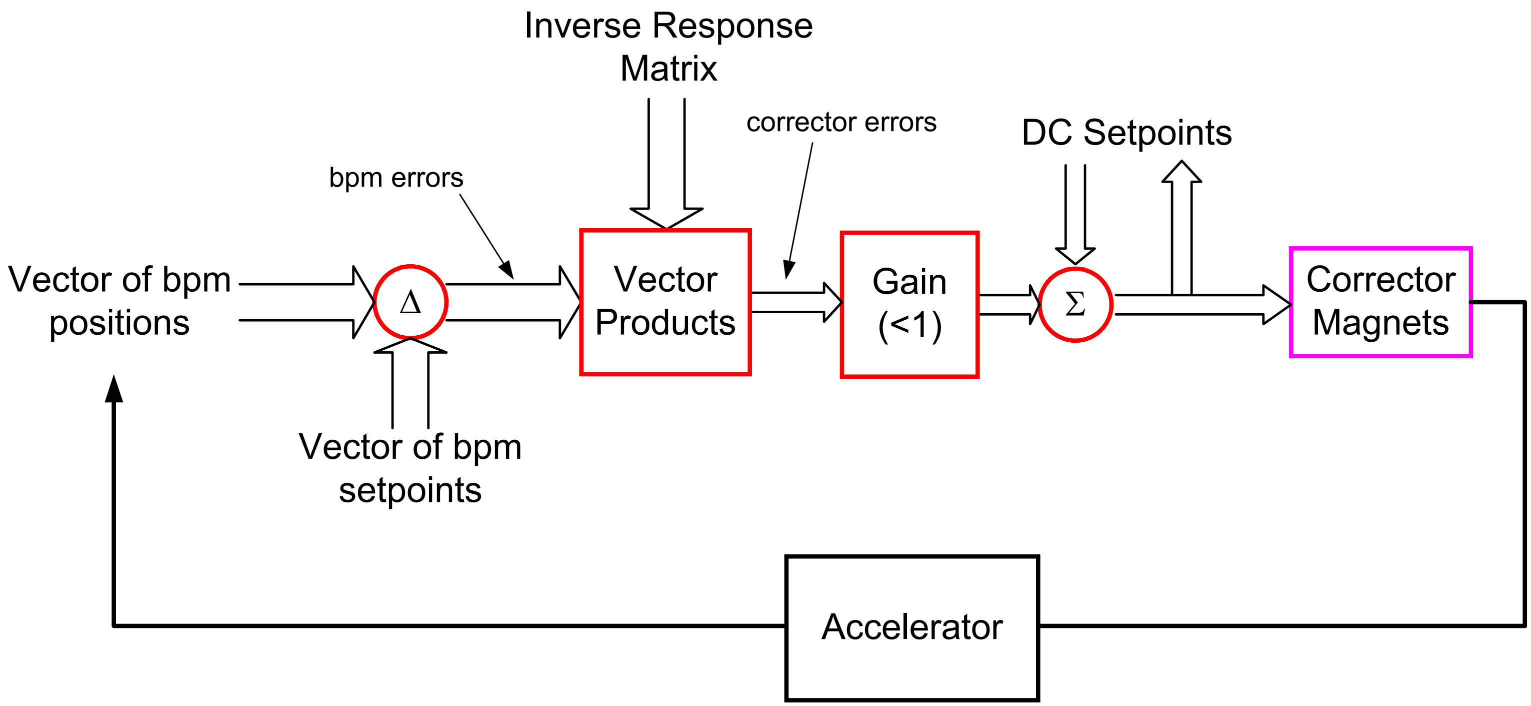
Where we've got to…

- Local, and global orbit correction algorithms provide a mechanism to measure the electron and/or photon beam position, and determine a set of corrector strength changes that will fix positional errors according to some criteria.

- The simplest implementation of this is to simply run the corrections in a repetitive loop, applying some fraction of the full computed correction (eg 40%) on each step, so the algorithm doesn't run away.

- Provided the corrections are done at slow repetition rates, say every few seconds (ie slower than system dynamics), this will work well, and will be stable.

The global orbit correction process



Orbit correction to orbit feedback

- Provided the system dynamics are fast compared with the correction rate, this implementation will work fine.
- Orbit correction is typically done at intervals from a few seconds to 100's ms.
- DC steering is often done only on demand
  - APS uses global orbit correction system to hold orbit at particular location.
  - When steering at x-ray source is needed, suspend orbit correction, make local steering, take snap-shot of new orbit, restart orbit correction to hold at new absolute orbit.
- To go beyond correction at faster than a few Hz, we have to know something about system dynamics, and include them in the regulator design.